

River Rapids

Open Source Intelligence (OSINT) – A Practical Introduction

A Field Manual

Varin Khera

Anand R. Prasad

Suksit Kwanoran



River Publishers

**Open Source Intelligence (OSINT) –
A Practical Introduction**

A Field Manual

Published 2024 by River Publishers

River Publishers

Alsbjergvej 10, 9260 Gistrup, Denmark

www.riverpublishers.com

Distributed exclusively by Routledge

605 Third Avenue, New York, NY 10017, USA

4 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Open Source Intelligence (OSINT) – A Practical Introduction/by Varin Khera,
Anand R. Prasad, Suksit Kwanoran.

© 2024 River Publishers. All rights reserved. No part of this publication may be reproduced, stored in a retrieval systems, or transmitted in any form or by any means, mechanical, photocopying, recording or otherwise, without prior written permission of the publishers.

Routledge is an imprint of the Taylor & Francis Group, an informa
business

ISBN 978-87-7004-717-3 (paperback)

ISBN 978-87-7004-719-7 (online)

ISBN 978-87-7004-718-0 (ebook master)

A Publication in the River Publishers Series in Rapids

While every effort is made to provide dependable information, the publisher, authors, and editors cannot be held responsible for any errors or omissions.

Open Source Intelligence (OSINT) – A Practical Introduction

A Field Manual

Varin Khera

Cloudsec Asia, Thailand

Anand R. Prasad

Japan

Suksit Kwanoran

SecStrike Co., Ltd.; Chief Technology Officer (CTO), CloudSec Asia Co., Ltd.





Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Contents

Preface	ix
----------------	-----------

About the Authors	xi
--------------------------	-----------

1 Introduction to Threat	1
1.1 Defining Intelligence	1
1.2 Threat Intelligence	3
1.3 Threat Intelligence Life Cycle	3
1.4 TI Types and Purpose	6
1.5 Key Threat Intelligence Terminology	8
1.6 Challenges and Limitations Associated with Threat Intelligence	10
1.7 Realistic Approach to Implementing TI	11
1.8 Open Source Intelligence (OSINT)	12
1.9 Book Overview	13

2 Introduction to Open Source Intelligence (OSINT)	15
2.1 OSINT Definition	16
2.2 OSINT Types	16
2.3 OSINT Users	17
2.4 OSINT Challenges	20
2.5 Chapter Summary	20

3	Online Tracking and Behavioral Profiling	23
3.1	IP Address	24
3.2	Cookies	25
3.3	ETag	27
3.4	Browser Fingerprinting	28
3.5	Chapter Summary	30
4	Hiding Your Traces When Conducting Online Investigations	33
4.1	Protect your Operating System	34
4.2	Secure Online Browsing	36
4.3	Countermeasures Against Online Tracking Techniques	37
4.4	Chapter Summary	41
5	Open Source Intelligence (OSINT): A Practical Example	43
5.1	Technical Investigation of a Website	43
5.2	Analytics and Tracking	45
5.3	Website History	46
5.4	Subdomain Discovery	47
5.5	Type and Versions of IT Infrastructure of the Target Company	49
5.6	Harvest Digital Files Hosted on Domains	50
5.7	Information Contained in File Metadata	51
5.8	Tools to Retrieve Digital File Metadata	53
5.9	Chapter Summary	54
6	Using AI in OSINT Research	57
6.1	Data Collection and Scraping	57
6.2	Analysis of Unstructured Text Data	58
6.3	Analysis of Multimedia Files (Images and Videos)	60
6.4	Content Summarization	60
6.5	Social Network Analysis	61

6.6 Chapter Summary	62
7 Social Media Intelligence (SOCMINT)	63
7.1 Privacy Issues In SOCMINT	65
7.2 OSINT Roadmap for Investigating Social Media Platforms	66
7.3 Chapter Summary	72
8 The Web Layers: Introduction to Surface, Deep and Darknet	75
8.1 Surface Web	76
8.2 Deep Web	77
8.3 Darknet	78
8.4 Chapter Summary	79
9 Darkweb and Internet Anonymity: Exploring the Hidden Internet	81
9.1 TOR Network	82
9.2 Searching the TOR Network	85
9.3 Chapter Summary	86
10 Introduction to Digital Forensics	87
10.1 Digital Evidence	88
10.2 Digital Forensics Process	90
10.3 Digital Investigation Types	91
10.4 Digital Forensics Readiness	92
10.5 Chapter Summary	93
11 OSINT for Digital Forensics Investigations	95
11.1 OSINT to Collect Individual Intelligence	95
11.2 Investigating Social Networking Sites	100
11.3 Investigating a Digital File's Metadata	101
11.4 Searching for Leaked Credentials on the Darknet	104

11.5 Chapter Summary	106
12 Data Protection and Cybersecurity Laws for the Asia Pacific Region	107
12.1 Classifications of Personal Information	108
12.2 Singapore	110
12.3 Japan	110
12.4 Vietnam	111
12.5 China	111
12.6 Thailand	113
12.7 Chapter Summary	114
Index	115

Preface

In an increasingly interconnected world, the ability to navigate and harness the power of open source information is an invaluable skill. The domain of open source intelligence (OSINT) offers unprecedented opportunities for individuals and organizations to enhance their situational awareness, uncover hidden insights, and make informed decisions based on publicly available data. This book is designed as a comprehensive and practical guide to OSINT, providing readers with the foundational knowledge and advanced techniques required to excel in this field. The chapters are structured to offer a step-by-step approach, ensuring that both beginners and experienced practitioners can benefit from the material presented.

We begin our journey with an introduction to threat intelligence, laying the groundwork for understanding the broader context in which OSINT operates. This initial chapter discusses the life cycle of threat intelligence, different types of intelligence, and the key terminology essential for mastering this discipline. Recognizing the challenges and limitations associated with threat intelligence, we aim to provide a realistic approach to its implementation, highlighting the significance of OSINT as a crucial component.

As we delve deeper, we explore the intricacies of OSINT itself, from its definitions and types to its diverse user base, which includes law firms, private investigators, ethical hackers, and government agencies. Each chapter is meticulously crafted to cover specific aspects of OSINT, such as online tracking, behavioral profiling, and techniques for conducting anonymous investigations. These topics are essential for understanding how to protect one's digital footprint while effectively gathering intelligence.

The practical applications of OSINT are illustrated through real-world examples, guiding readers through technical investigations of websites, meta-data analysis, and the retrieval of digital files. We also examine the role of artificial intelligence in enhancing OSINT research, demonstrating how advanced tools can streamline data collection and analysis.

Social media intelligence (SOCMINT) is another critical area covered in this book, providing a roadmap for investigating social media platforms while addressing privacy concerns. Additionally, we explore the different layers of the web, including the surface web, deep web, and darknet, offering insights into how to navigate these complex environments safely and effectively.

Digital forensics is an integral part of modern investigations, and this book dedicates chapters to introducing digital forensics concepts and integrating OSINT techniques into forensic investigations. By understanding how to collect and analyze digital evidence, readers will be better equipped to support comprehensive investigative efforts.

Finally, we address the legal and regulatory landscape, particularly focusing on data protection and cybersecurity laws in the Asia-Pacific region. This knowledge is crucial for ensuring that OSINT activities are conducted within the bounds of the law, respecting privacy and ethical considerations.

This book is not just a theoretical guide, but a practical manual filled with actionable insights, case studies, and exercises. Each chapter concludes with summaries and further reading suggestions to deepen your understanding and keep you abreast of the latest developments in the field. Whether you are a cybersecurity professional, a researcher, a law enforcement officer, or simply someone with a keen interest in open source intelligence, this book aims to equip you with the tools and knowledge to excel. Welcome to the world of OSINT. Your journey to mastering open source intelligence starts here.

About the Authors

Dr. Khera is a distinguished cybersecurity executive with more than 20 years of experience, currently serving as the CEO of CloudSec Asia (CSA) based in Thailand. Throughout his career, he has been at the forefront of developing and implementing cutting-edge cybersecurity solutions. Dr. Khera's expertise spans a wide array of fields within cybersecurity, including threat intelligence, cloud security, and the application of artificial intelligence in enhancing cyber defenses.

Under his leadership, CloudSec Asia has emerged as a leader in the cybersecurity industry, known for its innovative approaches to protecting organizations against evolving threats. Dr. Khera's profound knowledge and strategic insights in architecting security operations centers (SOCs) have been instrumental in mitigating cyber risks for high-profile global clients.

Prior to founding CSA, Dr. Khera held a key role as head of cybersecurity presales at Nokia, where he collaborated extensively with major telecom providers and government entities across the Asia Pacific region. He holds a Doctor of Information Technology (DIT) from Murdoch University, a Post-graduate Certificate in Network Computing from Monash University, and a Certificate of Executive Leadership from E-Cornell University.

Dr. Khera's contributions to the cybersecurity field have been widely recognized, including his receipt of the prestigious Asia Pacific Information Security Leadership Awards (ISLA) for excellence in IT Security Practitioner leadership. His commitment to advancing cybersecurity practices continues to shape the industry landscape globally.

Dr. Anand R. Prasad is a Partner at Deloitte Tohmatsu Cyber. He has also served as Board of Director at Digital Nasional Berhad. Prior to that, among other things, he was Founder & CEO of wenovator, acquired by DTCY, advisor to NTT DOCOMO and CISO, board member of Rakuten Mobile. Anand led

the standardization of 5G security as Chairman of 3GPP SA3. He is advisor to several organizations, an innovator with 50+ patents, a recognized keynote speaker and a prolific writer with 6 books and 50+ publications. He is a Fellow of IET & IETE, Editor of Cyber Security Magazine and Editor-in-Chief Journal of ICT Standardization by River Publishers.

Suksit Kwanoran is a seasoned expert in the field of cybersecurity and cloud computing, with over 17 years of extensive experience. He currently serves as the Managing Director of SecStrike Co., Ltd., where he leads the development of innovative security solutions and manages a high-performance team providing comprehensive security services. Additionally, he holds the position of Chief Technology Officer (CTO) at CloudSec Asia Co., Ltd., where he has significantly contributed to the company's rapid growth by integrating various technologies and expanding its range of services.

Suksit has a strong educational background, holding a Master's degree in Network Engineering and a Bachelor's degree in Electrical Engineering from Mahanakorn University of Technology. His career includes significant roles such as Senior Instructor at Network Training Center and IT Manager at Amata Corporation PCL. He is also a certified professional with numerous credentials including CompTIA CASP, CySA+, Pentest+, Security+, Cisco Certified System Instructor, CEH, ISO 27001 Lead Auditor, CCNA, and AWS Certified Solution Architect.

CHAPTER

1

Introduction to Threat

Digitalization is rushing to occupy all aspects of life. For businesses, digital transformation is no longer a luxury. Leveraging the latest technological advancement is crucial to achieve competitive advantages and survive in today's complex threat landscape.

This book will provide a practical introduction to open source intelligence (OSINT). Throughout this book, we will offer numerous technical tips and practical experiments to show you how to leverage free online data through a plethora of tools and online services to acquire intelligence from publicly available repositories.

Before we start our journey to learn OSINT, it is critical to understand the concept of cyber threat intelligence (CTI) and how OSINT fits into the CTI lifecycle to provide actionable intelligence to mitigate cyber threats originating from cyberspace.

In this introductory chapter, we start by taking a step back with a discussion on threat intelligence. With that base, we will discuss how to leverage OSINT tools and techniques in the following chapters.

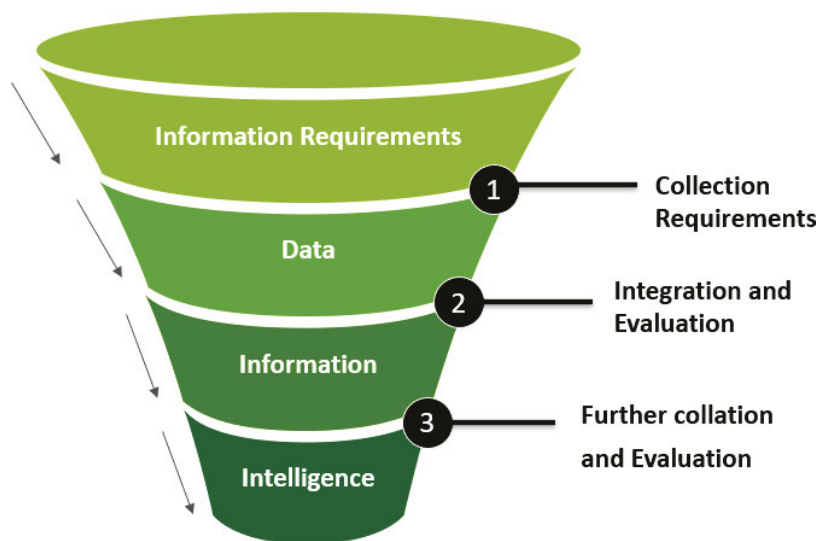
1.1 Defining Intelligence

Everything starts with data, thus we will start our discussion here from data, going on to information that leads to intelligence.

We are all well aware of the large amount of digital data available today due to the progress in digitalization and connectivity. In the cybersecurity context, data can come from various sources, such as network traffic logs, security event logs, malware samples, threat reports, OSINT (e.g., social media posts, news and blogs, websites), data collected by company operations centers – such as interactions with clients and server logs and more. All this data is unrefined and thus unusable as is.

Refining raw data, organizing and aggregating it gives us information that has context and is understandable. Processing information together with logic results in something that can be used to act upon, i.e., leading to informed decisions; this is called intelligence (see Figure 1.1).

Figure 1.1: Intelligence from data.



For example, a security operations center (SOC) collects a vast amount of raw data from various network devices (e.g., firewalls, intrusion detection systems, intrusion prevention systems, networking devices such as routers and Wi-Fi access points). This data might include timestamps, usernames, server names, and so on. While each piece of data has limited individual value for a security analyst, combining them can reveal valuable insights. When you combine different data elements, like a specific user accessing a server at an unusual time (or outside business working hours), you gain richer context. By analyzing and correlating this data over time, security analysts can better

understand user standard behavior patterns. This, in turn, allows them to take informed actions, such as requesting additional authentication steps when observed behavior deviates significantly from established patterns. This process of transforming raw data into actionable insights is what we refer to as threat intelligence.

1.2 Threat Intelligence

An example towards the end of the previous section has already taken us to threat intelligence (TI), which is at times also called cyber threat intelligence (CTI). In cybersecurity, TI plays a critical role that enables informed decision-making.

TI is based on knowledge and understanding of the given threat that usually constitutes information such as the threat actor's strategies, capabilities, limitations, relations and several other aspects. With that, TI can be used for several purposes, such as predicting future attacks, prioritizing remediation and plan effective defense solutions among others.

TI can be acquired from various sources, such as OSINT sources, closed-source intelligence (e.g., commercial threat feeds, industry-sharing groups), and internal data sources (e.g., servers logs, network traffic, incident reports, antivirus and antimalware programs logs).

TI forms the core for several security activities, some of these are listed below. In the following sections we give more detail regarding TI.

- Threat hunting and proactive defense – using TI to proactively look for threats that are not identified and using that knowledge to remediate potential issues thus proactive defence.
- Vulnerability management – prioritizing vulnerabilities based on their severity.
- Incident response and digital forensic analysis.
- Risk assessment and management – obtaining information about the current threat landscape and the motivations of potential adversaries.
- Security control optimization and resource allocation – for instance, TI provides specific details about the TTPs utilized by attackers in addition to justifying the resource allocation by knowing the types of threats an organization could be subject to

1.3 Threat Intelligence Life Cycle

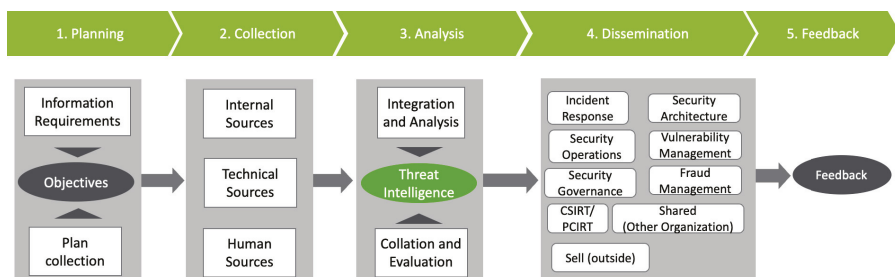
The threat intelligence cycle (also known as the intelligence cycle or the intelligence process) is a systematic approach used in the field of threat intelligence

to plan the gathering, analysis, and dissemination of intelligence related to cyber threats. This cycle is not a closed-circle process, instead, it is a continuous process to ensure that your organization's security team and top management are aware of the latest cyber threats that can affect their work activities.

The goal of TI depends significantly on the organization's objectives, which are determined by its risk profile and priorities. Data for TI is gathered from a plethora of internal and external sources that help identify potential threats against the organization so defensive measures can be taken promptly to avoid exploiting any weaknesses from threat actors.

A typical threat intelligence process is composed of (1) planning, (2) collection, (3) analysis, (4) dissemination, and (5) feedback; as detailed in following sub-sections. A holistic picture of TI is depicted in Figure 1.2.

Figure 1.2: A holistic view of threat intelligence.



1.3.1 Planning

In this phase, the work focuses on identifying the intelligence requirements and priorities based on the organization:

- Risk profile: As information security professionals we know that risk management forms the basis for any activity where the risk profile forms the base. This includes identifying the assets that should be protected along with the different types of threats targeting them.
- The type of IT assets and processes it needs to protect: This includes physical documents (paper documents), servers, storage devices containing customer data, financial records, intellectual property, customer relationships, brand reputation, and intellectual property. The business process side that should be protected includes order processing, online payments, and customer service interactions to name a few.

For example, a financial institution will prioritize gathering intelligence on threat actor groups who are known to target the banking industry or who have already targeted this organization in the past.

1.3.2 Collection

In the collection phase, we gather data from various internal and external sources to support our intelligence requirements. Some intelligence sources include:

- Checking logs of installed security devices across your IT environment, such as firewalls, servers, intrusion detection systems (IDS), intrusion prevention systems (IPS) and other networking devices.
- Getting feeds from free TI sources, such as AlienVault Open Threat Exchange (OTX) (<https://otx.alienvault.com>) and Phishtank (<https://www.phishtank.com>), in addition to free TI tools for visualizing threat data, such as the OpenCTI project (<https://github.com/OpenCTI-Platform/opencti>).
- TI commercial collection platforms are private companies that provide TI information on a subscription basis.
- OSINT data includes both online and offline sources, as we are going to see throughout this book.
- Monitoring darknet websites – such as discussion forums and chatrooms in the TOR darknet.

1.3.3 Processing

In this phase, security analysts convert raw data into usable information that can be used in a particular context – structured and unstructured data associated consideration comes here. For example, getting a list of malicious IP addresses from some TI feeds and incorporating them into an organization's security information and event management (SIEM) solution.

It is worth noting that processing can be performed by humans or automatically by machines.

1.3.4 Analysis

This phase commonly requires some human intervention. It works by interpreting collected data and converting it into actionable intelligence. Here are some examples of what might happen in this phase:

- Malware analysis: Security analysts may analyze the signature and behavior of newly discovered malware to identify its capabilities and who is behind it (threat actor group).
- Analysis of the indicators of compromise (IoCs) discovered during a recent cyber incident to gain more insight into the nature of the attack and who is possibly behind it.

1.3.5 Dissemination

In this phase, we turn the actionable intelligence into a final report and offer it to stakeholders (e.g., security operations, incident response, risk management, executive leadership) or other interested parties within or outside our organization. The goal is to use the collected intelligence to take specific actions or make more informed decisions.

The final report can be delivered via secure email, in a face-to-face meeting or using an organization's secure collaboration dashboard.

1.3.6 Feedback and improvement

We collect feedback from the recipients of the intelligence report to identify gaps and mitigate them in the next gathering activities. The feedback can be collected using different communication channels, such as secure portals and encrypted emails. It is critical to securely exchange threat intelligence information in compliance with implemented regulations to avoid breaching copyright or data privacy regulations.

Staying current on the latest threat landscape is critical, and it may require changing the tools or sources when gathering future intelligence.

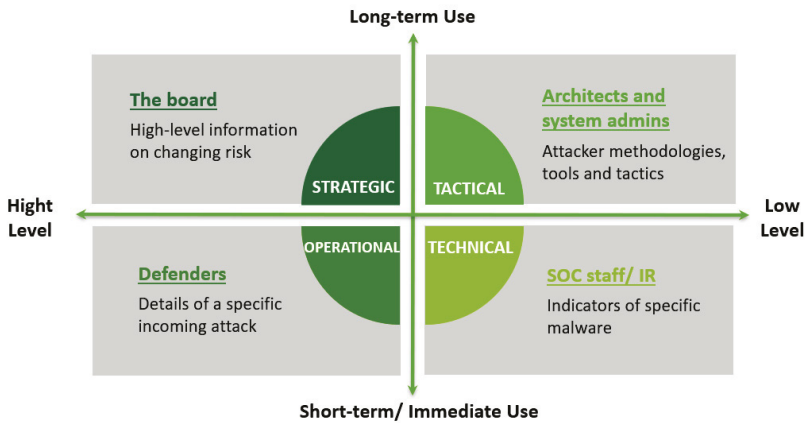
1.4 TI Types and Purpose

In general, TI can be divided into four different types: strategic, tactical, operational and technical, as explained below and depicted in Figure 1.3.

1.4.1 Strategic

This kind of information is usually high-level but still to the level that helps take strategic decisions at the highest level of an organization. Examples of strategic intelligence include the following:

Figure 1.3: Types of threat intelligence.



- Assessments of emerging cyber threats targeting the industry sector – such as ransomware attacks, supply chain attacks, and artificial intelligence (AI) attacks which are expected to increase in the coming years.
- Analysis of geopolitical trends impacting cybersecurity – for example, a cyberattack against Twain may disrupt the supply chain of EU and USA companies relying on Twainian semiconductor manufacture.
- Evaluations of regulatory compliance requirements impacting organizations' cybersecurity – for example, in the US, the Health Insurance Portability and Accountability Act (HIPPA) imposes robust data security for patient information. Similarly, PCI DSS (Payment Card Industry Data Security Standard) focuses on protecting credit card data.

1.4.2 Tactical

These are detailed TI regarding attackers. Such TI can help architecting solutions and thus should contain tactics, techniques, and procedures (TTPs). Examples of tactical intelligence include the following:

- Technical reports detailing specific malware families and their methods of propagation. Antivirus vendors can publish technical security reports detailing their discovery of new malware types. National cybersecurity agencies also publish such information.
- Analysis of recent phishing campaigns targeting employees. For example, a company can teach employees how malicious campaigns, such as Fake Two-Factor Authentication requests, work to steal target users' credentials.

- Assessments of exploit kits used in recent cyberattacks – for example, a company can analyze the recent cyberattack against its IT infrastructure and the type of vulnerabilities used to execute them. This helps security professionals understand attackers' methods and develop better defenses.

1.4.3 Operational

These kinds of TI help make operational decisions by developing appropriate operational solutions. The TI should contain specific information regarding the attacker, motivation, timing. Examples of operational intelligence include the following:

- Threat actor profiles outlining their motivations and objectives – for example, financially motivated cybercriminals will leverage ransomware attack vector to extort money from their victims. While Hactivist groups work to disrupt IT system operations using attacks such as distributed denial of service (DDoS) or to spread their political or ideological message via website defacement attacks.
- Timelines of recent cyber incidents affecting the organization's industry – for instance, by analyzing recent cyber incidents targeting similar companies, an organization can identify emerging threats and work to counter them. This knowledge is also beneficial to understanding the attack techniques employed by threat actors so our organization's security team can identify and block similar attempts in the future.

1.4.4 Technical

Technical intelligence (TECHINT) analyzes the technical indicators to understand adversaries' methods and proactively defend your IT systems. Examples of such indicators include:

- Malicious IP addresses – your security team may have a list of malicious IP addresses used by particular threat actors. They can add these IPs to the firewall blocked list to prevent your compromised devices from interacting with the hacker's command and control servers.
- Hashes of newly discovered malware can be added to the anti-malware or firewall solution to detect and prevent them from accessing the organization's IT environment.

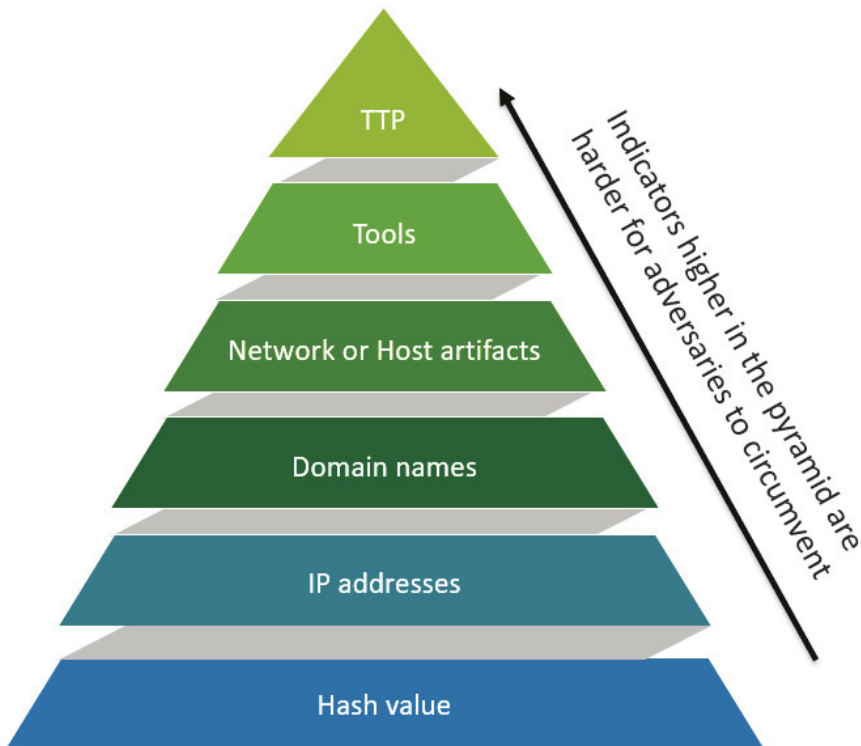
1.5 Key Threat Intelligence Terminology

Information used to identify potential threats or malicious activity is called indicators of compromise (IOCs). IOCs can vary in specificity and complexity,

ranging from simple to highly detailed. As the level of detail increases, gathering the IOC becomes more challenging (this is also known as the pyramid of pain, Figure 1.4):

- Hash values: This is the simplest level of information that can also be easily modified. Hashes are numerical values calculated from files or data, often used to identify malware samples or other digital artifacts uniquely.
- IP address: The IP address of the attacker can provide information about their geographical location, which is better than just the hash value; still, the IP addresses can also be modified using a virtual private network (VPN) or the TOR web browser.
- Domain names: The attacker's domain names that participated in the attack can provide more information about their infrastructure, such as their command-and-control servers or phishing sites.

Figure 1.4: Pyramid of pain.



- Network or host artifacts: The next level of information that can help distinguish an attacker from other. This will contain all kinds of digital artifacts left on compromised systems or networks, such as log entries, network traffic patterns, or system configuration changes.
- Tools: These are the software programs the adversaries use to perform cyberattacks such as exploit-kits and other hacking tools such as Mimikatz (<https://github.com/gentilkiwi/mimikatz>).
- Tactics, techniques, and procedures (TTP): This is detailed information with step-by-step information regarding how a given adversary performs a particular cyberattack.

The kill chain (see Figure 1.5) starts from reconnaissance and goes to weaponization, delivery, exploitation, installation, command and control, and finally, action on the objective. Until exploitation, one can be proactively defending, but from the exploitation state onwards, there is nothing other than the response phase and recovery.

Figure 1.5: Kill chain.



Obviously, one of the important things is to share TI among various parties. This is done by using structured threat information expression (STIX), cyber observable expression (CybOX) and trusted automated exchange of indicator information (TAXII).

1.6 Challenges and Limitations Associated with Threat Intelligence

Threat intelligence has become a valuable tool in security teams' arsenal to mitigate cyber-attacks before they reach your organization's IT environment. However, it still comes with its own set of challenges. Below are the main challenges organizations face when leveraging TI:

- Data volume: The world is digitalizing rapidly, and security teams will find themselves increasingly overwhelmed by the massive volume of data that they need to analyze. This makes their work harder as the volume of false positive alerts, outdated intelligence, and disinformation increases.

- Resources problem: Small and medium-sized businesses may not have the necessary resources (e.g., money, time, and expertise) to analysis the massive volume of threat intelligence data taken from different sources. This may lead to late responses or missing important alerts. Outsourcing this task is an option for limited budget organizations.
- Costs: TI can be acquired form free sources, such as AbuseIPDB (<https://www.abuseipdb.com>) and Binary Defense Systems Artillery Threat Intelligence Feed (<https://www.binarydefense.com/banlist.txt>) However, getting actionable threat intelligence from premium sources can be expensive and not affordable for all businesses.
- Integration issues: Integration of TI data with currently deployed security solutions and processes are daunting and may not always be feasible.

1.7 Realistic Approach to Implementing TI

As we already noted, TI comes with limitations; however, we can still overcome most of them by following best practices as realistic approach towards implementing TI in an organization:

- Source: It is better to focus on acquiring quality TI data rather than quantity. For example, obtaining intelligence from reputable sources, such as industry-specific threat-sharing communities, trusted commercial feeds, or intelligence from law enforcement agencies is far better than using free resources only.
- Relevance: The TI data should be relevant to your organization or industry. For example, some businesses are considered more lucrative targets for ransomware groups, while other companies could be more targeted for data theft or espionage.
- Training: User training is essential. For instance, your security team should possess valuable skills such as critical thinking, data analysis, threat modeling, and the ability to interpret and contextualize intelligence within the organization's IT environment. Continuous training and knowledge sharing are critical for security analysts' personnel as adversaries' attack techniques continue to evolve.
- Strategy: TI should be integrated into the overall cybersecurity strategy of the organization. For example, TI should be directly integrated with an organization's incident response and risk management strategy.
- Reliability: TI tools can be enhanced with machine learning (ML) and artificial intelligence (AI) capabilities. However, we should avoid over-reliance on automated solutions because of their susceptibility to false positive alerts.
- Continuous improvement: Finally, continuous monitoring and improvement are essential to stay ahead of emerging threats. Your security team must continuously measure the effectiveness of acquired intelligence and work to enhance the process of intelligence gathering, analysis, and dissemination.

1.8 Open Source Intelligence (OSINT)

OSINT, short for open source intelligence, derives its value from publicly available sources and is highly used for security purposes. It has become an essential and cost-effective component in sourcing threat intelligence from publicly available information. As we are going to see later in this book, OSINT can be acquired from a variety of sources, including the internet, newspapers, books, magazines, gray literature, and publicly available datasets – such as government databases; see Figure 1.6. However, a notable challenge lies in transforming the massive volume of unstructured, extensive datasets into actionable intelligence that can be used to protect organizations' IT assets from cyber threats. OSINT serves numerous use cases and is the focus of this book, explained further in the following chapters.

Figure 1.6: Open source intelligence (OSINT).



1.9 Book Overview

This chapter introduced the concept of cyber threat intelligence (CTI) with relevant background information regarding TI. With that background, in the following chapters we dive deeper into OSINT, a quick overview is given below:

Chapter 2: We introduce open source intelligence (OSINT), a powerful tool for gathering information from publicly available sources.

Chapter 3: We delve into online tracking, exploring how your digital footprint is collected and used.

Chapter 4: We discuss online privacy tips and best practices. This knowledge is critical for OSINT researchers to conceal their traces when performing online investigations.

Chapter 5: We put learning into action with a practical OSINT example.

Chapter 6: Using artificial intelligence (AI) and machine learning (ML) to support OSINT investigations.

Chapter 7: In social media intelligence, we will discuss using a general methodology to search social media networks.

Chapter 8: We differentiate between the surface web, deep web, and dark web, clarifying their unique characteristics.

Chapter 9: We shed light on the dark web, exploring its uses and potential risks.

Chapter 10: We introduce the field of digital forensics and its methods for recovering and analyzing digital evidence.

Chapter 11: We examine how OSINT plays a crucial role in digital forensics investigations.

Chapter 12: We wrap up the book by exploring the legal considerations surrounding the use of the information and techniques presented.

Further Reading

1. MWR InfoSecurity, "Threat Intelligence: Collecting, Analysing, Evaluating", <https://www.foo.be/docs/informations-sharing/Threat-Intelligence-Whitepaper.pdf> Accessed 2024-04-26
2. Medium (Ensar Seker), "Deciphering the Digital Battlefield: Navigating the Complex World of Cyber Threat Intelligence", <https://ensarseker1.medium.com/deciphering-the-digital-battlefield-navigating-the-complex-world-of-cyber-threat-intelligence-17d90b937278> Accessed 2024-04-26
3. National Cyber Security Centre, "An Introduction to Threat Intelligence", <https://www.ncsc.gov.uk/files/An-introduction-to-threat-intelligence.pdf> Accessed 2024-04-26

CHAPTER

2

Introduction to Open Source Intelligence (OSINT)

The internet has changed everything around us, from education, healthcare, and government interactions reaching to social communication, which receives the greatest impact. The internet has redefined how people communicate with each other and revolutionized how corporations do business. Nowadays, the majority of world communications happen in what is known as cyberspace.

According to *cybersecurity ventures*¹, by 2030, 90% of the human population aged 6 years and older will be online; this means more than 7.5 billion internet users. People now use the internet to purchase goods and services, entertainment, connect with other people, share information and files, in addition to using social networking websites to communicate with friends and family members without any geographical barriers.

As the world continues to digitalize, digital societies will produce a huge amount of digital data generated from people and business interactions in cyberspace. Exploiting this info in the right direction will open up numerous opportunities for public and business organizations to increase profits and operate more efficiently in the new information age.

Open source intelligence (OSINT) refers to all information that can be found publicly –mostly via the internet – without breaching any copyright or privacy laws. Under this definition, a wide array of sources can be considered a part of OSINT. For instance, information posted publicly on social media websites, posts on discussion forums and group chats, unprotected websites directories (such as

¹Cybersecurityventures, “Humans On The Internet Will Triple From 2015 To 2022 And Hit 6 Billion” <https://cybersecurityventures.com/how-many-internet-users-will-the-world-have-in-2022-and-in-2030/> Accessed 2024-05-01

open FTP servers) and any piece of information that can be found by searching online. Keep in mind that most OSINT resources cannot be found using regular search engines such as Google or Yahoo!, as many resources are buried deep in the web and darknet; such resources constitute more than 96% of the web content.

In this chapter, we will shed light on the term OSINT, discover its types and actors interested in OSINT gathering, and explore OSINT's benefits in today's digital age.

2.1 OSINT Definition

As we have already mentioned, OSINT refers to all the information available for public consumption, including online and offline resources. You may wonder if this information needs to be free to be considered a part of OSINT resources. The answer is no. For example, the information in scientific papers, books, and magazines must be purchased first to disseminate it in your OSINT gathering activity.

The US Department of Defense (DoD) defines OSINT as follows:

“Open source intelligence (OSINT) is an intelligence that is produced from publicly available information and is collected, exploited, and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement.”

2.2 OSINT Types

OSINT can be classified according to where the publicly available information is found, such as:

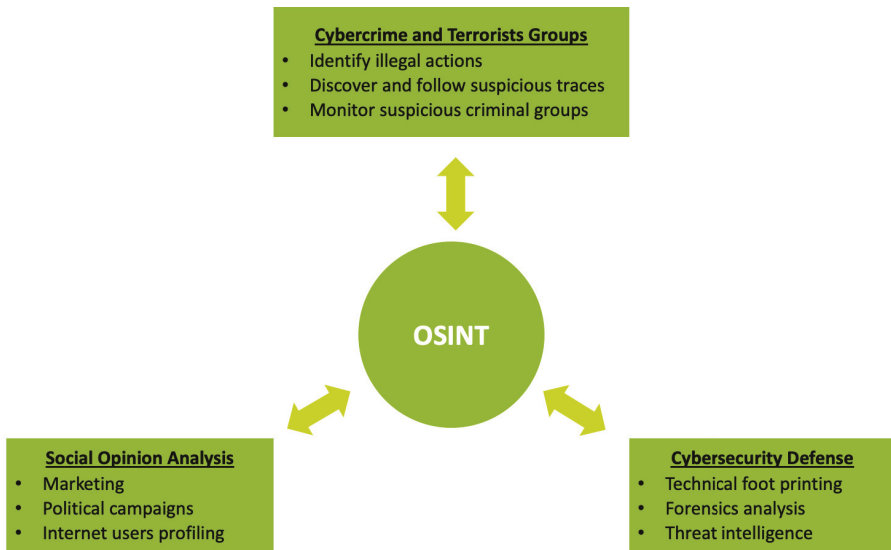
1. The internet is the main place where OSINT resources are found. Indeed, many researchers differentiate between the online OSINT resources and the offline ones by using the term “cyber OSINT” to refer to internet resources exclusively. Internet resources include the following and more: blogs, social media websites, digital files (photo, videos, sound) and their metadata, technical foot printing of websites, webcams, deep web (government records, weather records, vital records, criminal's records, tax and property records), darknet resources, data leak websites, IP addresses, and anything published online publicly.
2. Traditional media channels such as TV, radio, newspapers, and magazines.

3. Academic publications include dissertations, research papers, specialized journals, and books.
4. Corporate papers include company profiles, conference proceedings, annual reports, company news, employee profiles, and résumés.
5. Geospatial information includes online maps, commercial satellite images, geo-location information associated with social media posts, and transport (air, maritime, vehicles, and railway) tracking.

2.3 OSINT Users

Actors or users have varying motivation to gather OSINT; see Figure 2.1. In this section we discuss details regarding users and their interests in OSINT.

Figure 2.1: OSINT users and motivation.



2.3.1 Law Firms and private investigators

OSINT is used extensively by law firms to optimize their litigation by discovering information found on social media sites and other online places to uncover biases and acquire important information about the individual's or organization in question. The information acquired from public sources can be beneficial in the following cases:

- Discrimination and sexual harassment lawsuits
- Wrongful termination, disability, and hostile work environment claims
- Intellectual property violation cases.

2.3.2 Ethical hackers

IT security professionals utilize OSINT search techniques and tools to discover weaknesses in friendly IT systems, so such vulnerabilities can be closed before threat actors discover them. Commonly found vulnerabilities include:

1. Accidental leaking of sensitive information on social media sites. For example, an unaware employee may post a personal photo taken in the server room showing the type of security devices used to secure a corporate network.
2. Open ports and insecure services running on servers and endpoint devices can be discovered when scanning the subject network for vulnerabilities using specialized tools.
3. Outdated operating system versions, software and any content management systems already in use.
4. Leaked information found on data leak repositories or across the darknet.

2.3.3 Business competitive analysis

As the internet becomes widely adopted in all life and business areas, corporations can utilize OSINT to gain great insight into current and future threats. For example, OSINT can be used to gain useful intelligence about competitors' marketing, business operations and expansion strategies, and their deals with other companies in addition to their future plans (e.g. expand to new markets, launch new products or services).

2.3.4 Law enforcement agencies

OSINT techniques help law enforcement officials improve their intelligence-gathering activities to protect citizens and businesses from cybercriminals. OSINT can also be utilized in this context to identify possible criminals – by examining their social media accounts and online behavior – before they commit their crime. For example, law enforcement can use a search algorithm to scan social media sites – and other online public sources sites – for terms like “shoot” or “kill” to stop possible criminals before conducting any crime.

2.3.5 Government agencies

Governments are the greatest consumers of OSINT; they need such info to predict future trends on a global level. Governments seek professional reports concerning any area of interest (political, health, economic or sports events, etc.) from specialized OSINT firms to help them in their decision-making process.

2.3.6 Individuals

Ordinary people use OSINT to check how much personal information is exposed about them online. This helps them discover and delete any unwanted information leaked publicly and prevent bad actors from exploiting such info to target them with customized attacks (e.g., social engineering attacks).

In general, all internet users use some OSINT search technique in one way or another. For example, when using Google to search for something or when using the search box on Facebook or Twitter to search for someone, you are utilizing OSINT to find this info.

2.3.7 Cybercriminals and terrorist organizations

On the bad side, cybercriminals and terrorists are using OSINT techniques in the same way good people use to find information about their targets. Threat actors use OSINT to examine possible targets, identify weaknesses in target computer networks, and use this intelligence to exploit the target.

OSINT is considered a valuable tool to assist in conducting social engineering attacks. The first phase of any penetration testing methodology begins with reconnaissance – in other words, with OSINT; see Figure 2.2.

Figure 2.2: General penetration testing methodology always begins with OSINT gathering (reconnaissance).



2.4 OSINT Challenges

Despite the numerous advantages of OSINT for various parties, OSINT gatherers will face many challenges when acquiring it. Here are the most common challenges associated with OSINT gathering:

1. **Data volume:** The massive volume of publicly available data can be overwhelming. For instance, OSINT researchers need to use different tactics and analyze tools to filter gathered data to identify which is relevant to the investigating case.
2. **Verification problems:** Verifying the accuracy and authenticity of information gathered from OSINT can be challenging. For instance, threat actors may spread disinformation to mislead investigations. OSINT researchers must cross-check information from multiple sources to identify biased and outdated information and remove it from their final report.
3. **Time consumption:** Collecting and analyzing information from publicly available sources can be time-consuming, especially when dealing with complex investigations.
4. **Incomplete information:** Information collected from OSINT might be incomplete, leading to difficulty drawing definitive conclusions or connecting different dots to form a solid result.
5. **Technical skills:** While some OSINT techniques require little technical expertise, leveraging advanced tools and automating data collection may require specific programming or scripting skills – such as using the Python and R programming languages.
6. **Ethical issues:** It's vital to ensure that your OSINT gathering activities adhere to legal and ethical boundaries. Privacy laws in many countries, such as the GDPR in the EU, prevent the collection and storage of EU citizens' personal information without following specific guidelines. We should also consider the copyright regulations when using grey literature in our investigations.
7. **Language barriers:** Many OSINT investigations require collecting data from various sources in other languages you don't understand; language barriers can hinder your ability to access and comprehend crucial data.
8. **Evolving landscape:** The digital landscape and the methods of collecting and sharing information constantly evolve. Staying updated on new tools, techniques, and legal considerations is essential for effective OSINT practice.
9. **Disinformation:** Threat actors may try to spread fake information online to confuse OSINT researchers. This will lead to misinterpretations and misleading investigations. Critical thinking and source evaluation are crucial for filtering out false information.

2.5 Chapter Summary

Pushed by the huge technological advancement and the wide prevalence of internet communication worldwide, OSINT has become a critical component of both public and private intelligence, supplying businesses, governments, and

individuals with a plethora of tools and techniques to gather intelligence from high-quality information to base your decisions according to it.

OSINT is beneficial for different scenarios, whether you are investigating for research, competitor intelligence, vulnerability assessment, threat analysis, or you are simply an individual who cares about their privacy and wants to discover what personal information is already – inadvertently – leaked about them, OSINT will give you the required tools to have access to some of the best available data in the world and mostly for free.

This chapter serves as an introduction to the OSINT subject; in the coming chapters, we will explore how to implement several OSINT techniques to gather intelligence using online public sources. However, before we begin the gathering process, we should first discuss how our activities can be traced online and the best methods to hide them when conducting online investigations. Both topics will be discussed in the coming two chapters.

Further Reading

1. Secjuice, "An Introduction To Open Source Intelligence (OSINT) Gathering", <https://www.secjuice.com/introduction-to-open-source-intelligence-osint> Accessed 2024-05-01
2. Sans, "What is Open Source Intelligence?", <https://www.sans.org/blog/what-is-open-source-intelligence> Accessed 2024-05-01
3. Recordedfuture, "What Is Open Source Intelligence and How Is it Used?", <https://www.recordedfuture.com/blog/open-source-intelligence-definition> Accessed 2024-05-01
4. Hassan, "Open Source Intelligence Methods and Tools: A Practical Guide to Online Intelligence", Apress, 2018, <https://www.amazon.com/Open-Source-Intelligence-Methods-Tools-ebook/dp/B07F5Y6P56> Accessed 2024-05-08



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER

3

Online Tracking and Behavioral Profiling

As the world continues to digitalize, the practice of web tracking has grown increasingly to an extent that threatens people's privacy. In today's digital age, anything you do online can be tracked and recorded in some way! Yes, you read it correctly, anything. And if you think activating the (private) incognito mode in your web browser will prevent others from tracking you online, I'm afraid you are completely wrong.

Different actors are interested in tracking and recording internet users' activities, such as online advertisers, government agencies, website owners (including social media platforms), search engines and internet service providers (ISP). Online advertisers are the leading group. They record internet users browsing activities to formulate a complete online "profile" for each connected user; later these profiles will be used to display tailored advertisements to internet users based on their online behavior.

Online tracking poses serious privacy concerns for the general public. For instance, sensitive information, such as financial and health information, is usually collected. In addition, anything an internet user asks when using search engines will also be recorded and added to their online profile. This tracking profile will uniquely distinguish an internet user whenever they go online. The general public thinks that formulating internet profiles for their browsing activities does not pose a privacy risk, as the collected data is anonymous and cannot be linked to their real-world personality. However, this is not always true. Online trackers can easily link the historical browsing data of any internet user to their real identity using various methods.

While tracking ordinary internet users has become a privacy concern for the general public, online investigators (such as digital forensics investigators and OSINT gatherers) who surf the internet to collect intelligence for a variety of reasons need to be very careful in preventing outside observers from tracking their online activities. The first prerequisite of any OSINT gathering task is to secure your digital footprint, i.e., conceal it, and to do that correctly. To do that you must know how others can track you online.

Trackers employ various technologies to track internet user's behavior while new techniques are constantly developing; many of them are hard-to-detect techniques that can track an internet user across various devices, e.g., laptops, smartphones, smartwatches and any Internet of Thing (IoT) device such as health devices. In the following sections we introduce the most common online tracking techniques in use.

3.1 IP Address

Whenever you connect to the internet, your computing device can be identified by a unique number called the internet protocol – or IP.

An IP address is a unique number that distinguishes your device when connecting to the internet. For instance, no two devices can have the same IP address on the same network. There are two versions of IP: IPv4 (32 bits long) has the format (192.168.1.1), and IPv6 (128 bits long) has the format (0:0:0:0:ffff:c0a8:101).

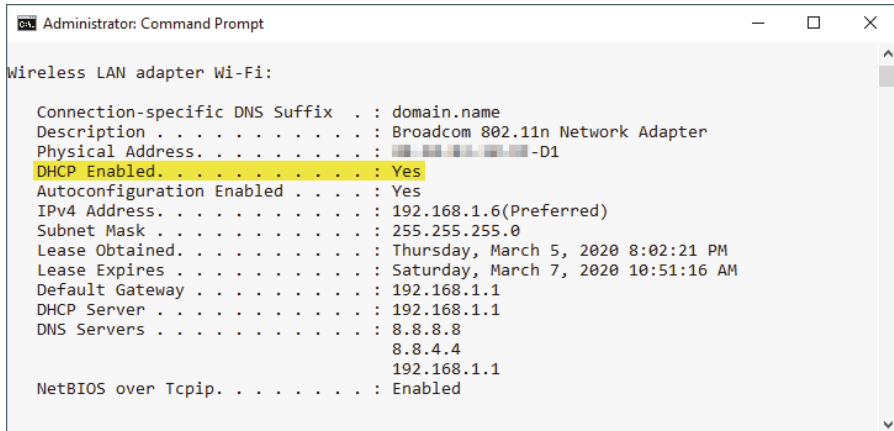
IP addresses come in two types: static and dynamic. A static IP address does not change and remains the same even after the user reboots their computing device or router. Static addresses are commonly used in email and file storage servers.

A dynamic IP address changes every time a user reboots their computing device; your network administrator or your ISP assigns it for a limited time. Dynamic addresses are commonly assigned using the dynamic host configuration protocol (DHCP), a network protocol used to dynamically assign an IP address for each device on the network. Most internet users use a dynamic IP address when connecting to the internet.

To know whether you are using a static or dynamic IP address, type `ipconfig /all` on your Windows command prompt, hit the Enter button, navigate to your current network connection, and search for DHCP. If the DHCP Enabled

parameter is set to Yes (see Figure 3.1), then you most likely have a dynamic IP address.

Figure 3.1: Checking connection settings under Windows OS.

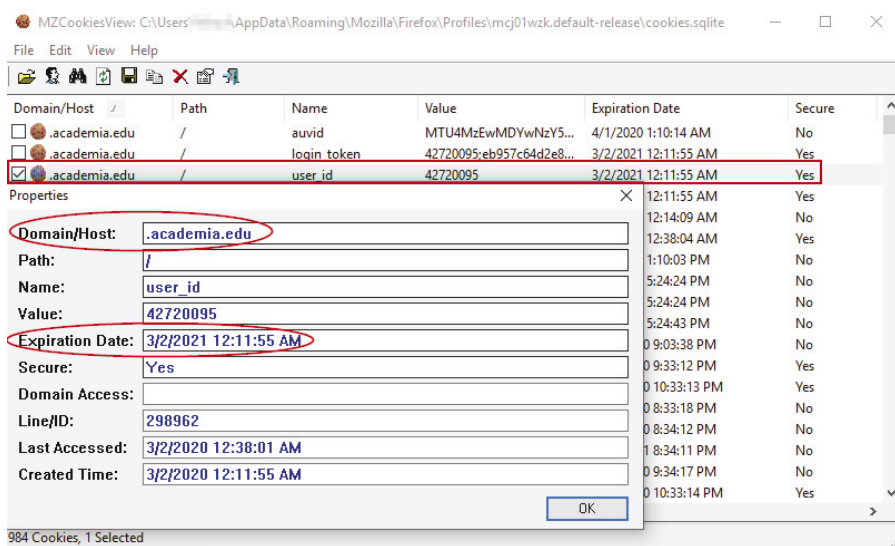


An IP address is the first option online trackers use to track internet users' browsing activities; however, we cannot consider an IP address as the sole unique identifier of internet users. For example, an IP number can be concealed using a virtual private connection (VPN) or anonymous networks such as the TOR network. This fact made online trackers use other techniques – in addition to the IP address – to recognize online users as we will see next.

3.2 Cookies

This is one of the oldest techniques used to track internet users. In its simplest form, a cookie is a small text file stored on a user web browser when visiting a website that uses cookies for the first time. A cookie was invented to remember users' preferences when visiting a website so website owners can personalize their contents – such as user location, language and theme preferences – when the user returns to the same site again. A cookie file can store different information according to its type; simple text cookie files contain the name of the URL (domain name) the cookie belongs to in addition to the cookie expiration date (see Figure 3.2).

Figure 3.2: Using MZCookiesView from nirsoft <https://www.nirsoft.net/utils/mzcv.html> to view all text cookies stored within Mozilla Firefox.



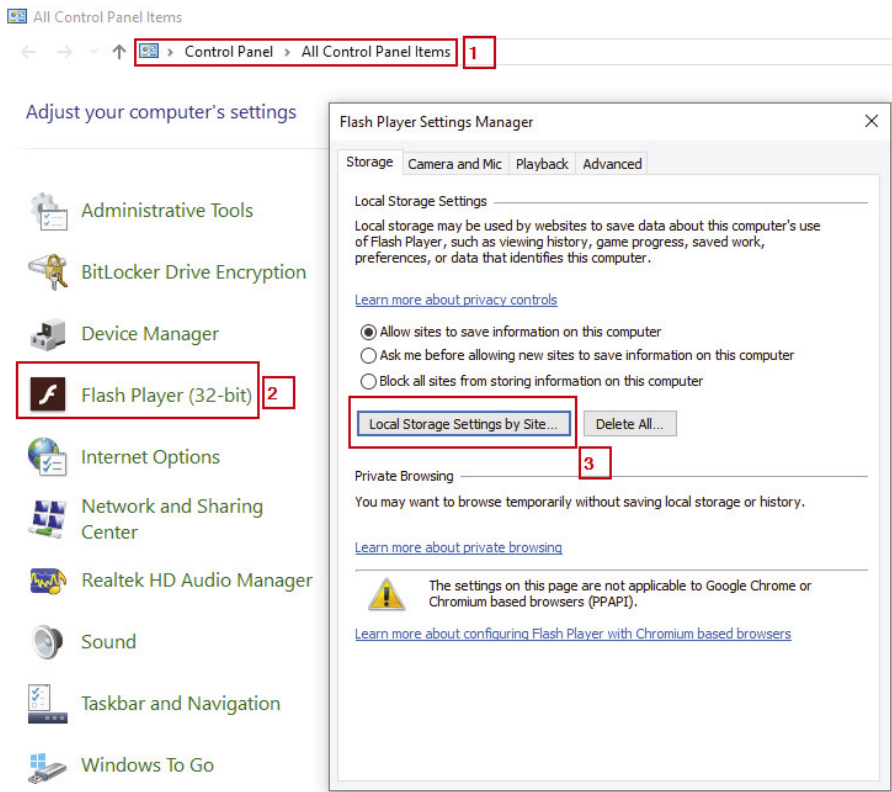
When the website you visit directly installs the cookie on your device, then this cookie is called a first-party cookie, while third-party cookies are installed by websites other than the one you are currently on. Third-party cookies are the ones that impose privacy risks for internet users because they can track users browsing history across multiple websites.

Cookies can be grouped according to their expiration date as:

Session cookies: This type has no expiration date; session cookies are deleted automatically when the user terminates the session or closes the web browser.

Persistent cookies: Usually used to store user's settings (language, theme, menu preferences) and to facilitate other useful functions such as authentication. A persistent cookie can live for a long time (until its expiration date). Many third-party advertisers set no expiration date for their cookies, making it live for good unless a user deletes it. The most known type of persistent cookies is Flash cookies. (Figure 3.3) demonstrate how to access all Flash cookies stored on your windows computer.

Social media platforms like Facebook use third-party cookies to track internet users. Facebook achieves this technically through its "Like" and "Share"

Figure 3.3: Viewing all Flash cookies stored on a user device under Windows OS (all versions).

buttons, already spread over the internet. Facebook includes a small tracking code (JavaScript) within its buttons. So, whenever a user visits a website that holds these buttons, Facebook will automatically record this action and begin tracking the user even though they do not own a Facebook account.

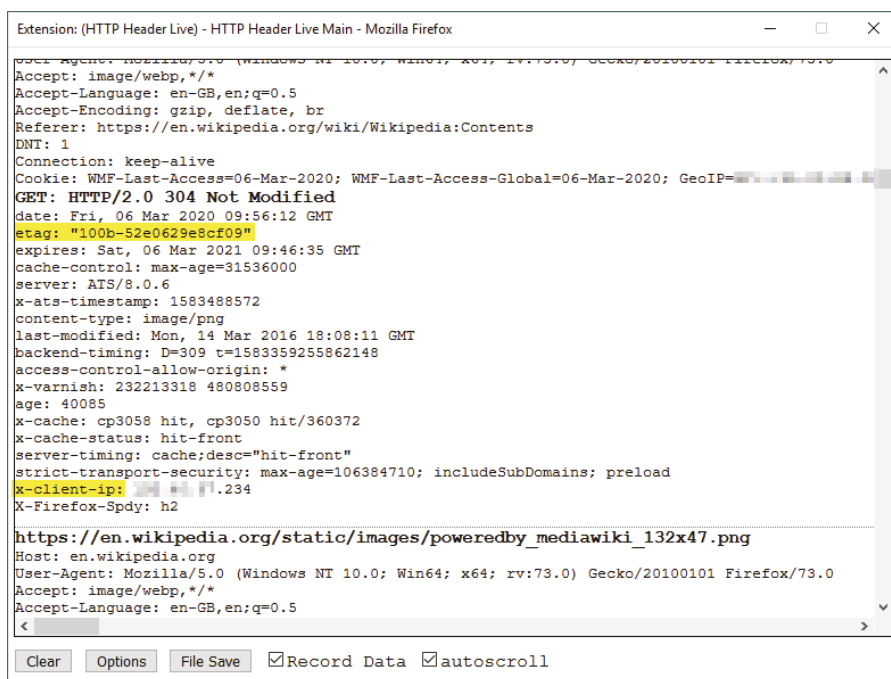
Online trackers commonly use cookies along with IP addresses to track internet users browsing history more accurately.

3.3 ETag

An entity-tag (ETag) is an HTTP mechanism that provides web cache validation to increase the internet surfing speed of end-users when visiting supported

websites. ETag works by comparing the resources (images, videos, and audio files) on the end-user machine with those on the visited website. If the version of the local resources stored on the end-user device is the same, then there is no need to download the resources again from the web server; see Figure 3.4.

Figure 3.4: Viewing Wikipedia website header info using Firefox browser extension HTTP Header Live (<https://addons.mozilla.org/en-US/firefox/addon/http-header-live>) Both ETag and IP address tracking is used in this example.



ETags can be exploited by online trackers by forcing the tracking server to send continuous ETags even though nothing changed on the webserver. This maintains an active connection (session) between the end-user device and the tracking server that lasts indefinitely without user knowledge.

3.4 Browser Fingerprinting

Also known as “digital fingerprinting”, in this type a user computing device is uniquely identified online using its technical specifications/settings.

Fingerprinting works by running a code (usually JavaScript, Java applet or Flash code) inside the user's web browser. Upon execution, this code will extract different technical information about the target web browser and device settings, such as:

- Screen resolution
- Processor type
- Memory
- Installed web browser add-ons
- Installed fonts
- Language settings
- Time zone
- Location
- Browser type
- Operating type/build number
- User behavior attributes, such as typing speed, mouse movements and browsing habits.

After the required information is gathered, a hash is made based on the collected information which is used to track users across the internet.

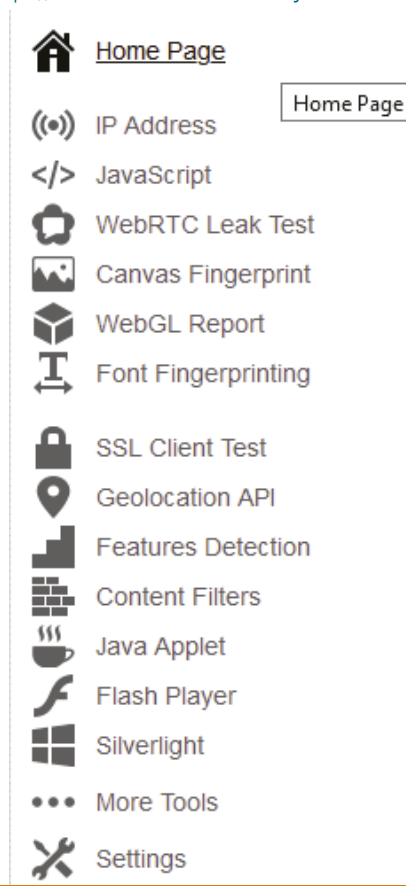
Device fingerprinting allows trackers to track internet users transparently without using cookies or IP addresses. Some people may think that the technical data collected from end-users' devices using this method is generic and cannot be used to distinguish a user's computing device among millions of connected devices. However, this assumption is not accurate. For instance, in a paper EFF in 2010², they found that most web browsers used in their experiment can be uniquely identified using this technique. Although the study is somehow old, it is still valid, especially with the continual development of fingerprinting technologies.

There are different online services to see your current device/browser fingerprinting. The following are the most popular ones:

1. PANOPTICCLICK (<https://panopticlick.eff.org>)
2. AmlUnique (<https://amiunique.org>)
3. Browserleaks (<https://browserleaks.com>) (see Figure 3.5)

²EFF, "How Unique Is Your Web Browser?" <https://panopticlick.eff.org/static/browser-uniqueness.pdf>, Accessed 2024-03-09.

Figure 3.5: Use <https://browserleaks.com> to check your current browser fingerprint



3.5 Chapter Summary

Online tracking is an important subject to understand for both cybersecurity professionals and end-users. For internet investigators, knowing how to conceal your online traces is very important before beginning your search, and to do that correctly, you should first understand how others can track your activities online.

In this chapter, we presented a high-level overview discussion of the most common web tracking technologies. In the next chapter, we will continue the discussion in the same context towards methods to prevent outsiders from

recording your browsing activities when conducting OSINT research using a plethora of tools and techniques.

Further Reading

1. Cookiebot, "What are tracking cookies and how do they work?", <https://www.cookiebot.com/en/tracking-cookies> Accessed 2024-05-08
2. USA Federal Trade Commission, <https://consumer.ftc.gov/articles/how-websites-and-apps-collect-and-use-your-information> Accessed 2024-05-08
3. Secjuice, "Tracking The OSINT Hunter" <https://www.secjuice.com/tracking-osint-hunters> Accessed 2024-05-08
4. GCF Global Learning, "Understanding browser tracking" <https://edu.gcfglobal.org/en/internet-safety/social-media-privacy-basics/1> Accessed 2024-05-08



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER

4

Hiding Your Traces When Conducting Online Investigations

Online privacy (also known as internet privacy) is a general term that refers to a wide array of techniques and technologies used to secure personal and confidential information when going online. Internet privacy has become a hot topic in today's digital age, and it is a growing concern for the general public. Global media channels announce daily news about internet privacy violation incidents that mostly come in the form of leaking confidential business or user's data for malicious purposes.

There is no sign that cybercriminal activity will slow down in the future; on the contrary, all estimates show that it will increase exponentially. For instance, *Cybersecurity Ventures* predicted that cybercrime will cost the world US\$9.5 trillion a year by 2024³, and a large portion of this number comes as direct or indirect losses caused by internet privacy breaches.

Aside from the concerns of the general public about the security of their personal information and communication online, online privacy is crucial for IT security professionals, especially OSINT gatherers, who need a high level of privacy and anonymity when conducting online investigations.

In this chapter, we will give a high-level overview on the subject of online privacy using practical tips. We will point to the many online resources and give technical advice on how to keep online communications secure and prevent cyber-attacks against computing devices.

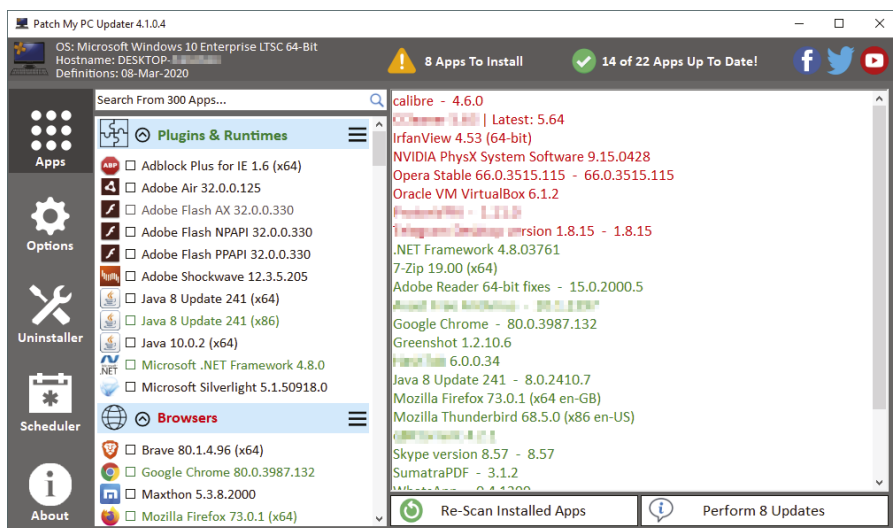
³Cybersecurity Ventures, "Cybercrime To Cost The World \$9.5 Trillion USD Annually In 2024" Accessed 2024-03-07 <https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016>

4.1 Protect your Operating System

The first thing you need to consider when securing your online privacy is to ensure your operating system (OS) is secure and does not suffer from any security problems or open holes that allow outside intruders to gain unauthorized access to your device. The following measures should be followed to secure your OS (we will use Windows as an example in some tips, as it is the most used OS on desktops and laptops worldwide).

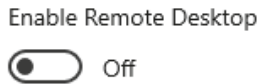
1. Install reliable antivirus software, dedicated anti-malware software, and personal firewall solutions. *Comodo* offers a free personal firewall solution (<https://personalfirewall.comodo.com>), and *Avast* has a free version of its antivirus solution (<https://www.avast.com/free-antivirus-download>).
2. Keep your OS up-to-date, and never use discontinued OS versions such as Windows 7 and XP.
3. Keep your installed applications up-to-date. You can use a software updater to do this. *Patch My PC* (<https://patchmypc.com/home-updater-download>) is an example, see screenshot in Figure 4.1.

Figure 4.1: Using Patch My PC Home Updater to automatically update installed applications.

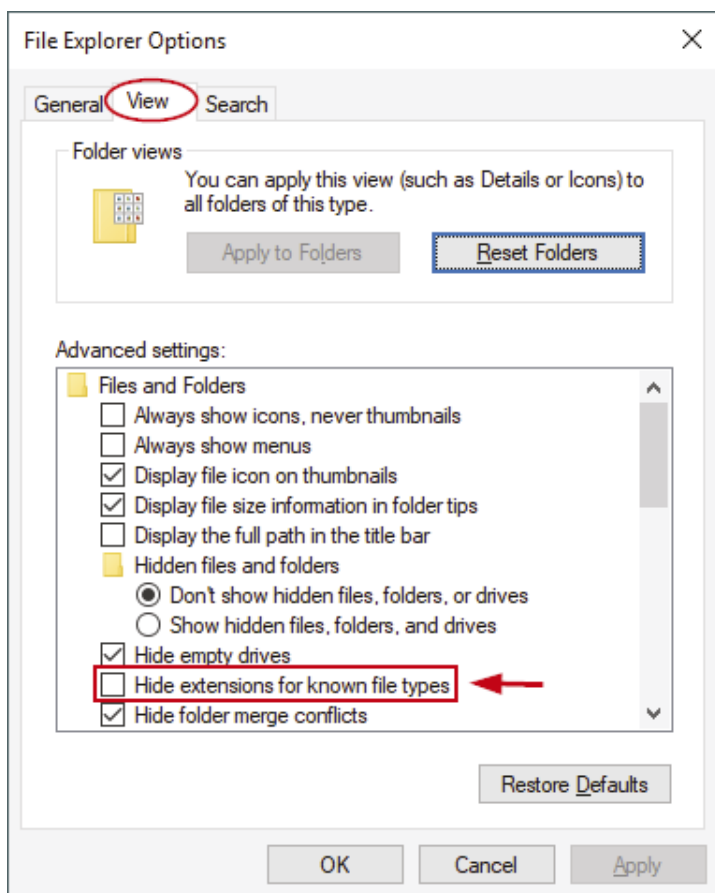


4. Use a less-privileged user account when conducting OSINT searches. Using an administrator account is unnecessary; instead, use a limited user account for your daily tasks. This will prevent many types of malware from infecting your device. To set up a restricted user account under Windows 10, check this detailed guide (<https://www.laptopmag.com/articles/limited-user-accounts-windows-10>).
5. Secure your OS login with a strong password.
6. Do not install pirated software from free file sharing or Torrent websites, and do not install internet programs unless you trust the program and the hosting website.
7. Disable remote desktop protocol (RDP) to prevent remote access to your device using this protocol. To disable RDP under Windows 10, follow these steps:
 - ✓ Press *Windows key + X* buttons to open the *Quick Access* menu, click on *System*
 - ✓ On the left-hand side, click on *Remote Desktop*, and turn it Off (see Figure 4.2)

Figure 4.2: Disable RDP under Windows 10.



8. Use full disk encryption to protect your computing hard drive. Make sure to encrypt your USB portable devices as well, especially if you are using them to store sensitive files. Windows users can activate the built-in encryption utility "BitLocker" which comes with most modern Windows versions. You can find a complete guide on how to use BitLocker: https://www.groovypost.com/howto/use-bitlocker-encryption-windows-10_
9. Do not run macros in MS Office files unless you trust the file's sender.
10. Show file extensions to recognize potentially malicious files. To view the file extension on Windows 10, go to *Control Panel* and select *File Explorer Options*, go to the *View* tab, and deselect the option *Hide extensions for known file types* (see Figure 4.3).
11. Cover your laptop camera and microphone to prevent hackers from recording video/audio if they compromise your device.

Figure 4.3: Show file extension under Windows 10.

4.2 Secure Online Browsing

Web browsers are your window to the Web, so you should configure your web browser to use tight privacy options. *Restoreprivacy* (<https://restoreprivacy.com/firefox-privacy>) offers a detailed guide for configuring Firefox browser to become more privacy-oriented. We focus on the Firefox browser because it is an open source browser. As we will see in a coming chapter, a modified version of Firefox is also used to access the TOR darknet, it is called the Tor browser.

Several privacy add-ons can help maintain a user's privacy online; the following are the most popular trusted add-ons for the Firefox browser:

1. Privacy Badger (<https://www.eff.org/privacybadger>): Blocks spying ads and invisible trackers.
2. HTTPS Everywhere (<https://www.eff.org/https-everywhere>): Encrypts communications with major websites, making your browsing more secure.
3. NoScript (<https://noscript.net>): Allows JavaScript, Java, Flash, and other plugins to be executed only by trusted web sites that are selected by the user.
4. uBlock Origin (<https://addons.mozilla.org/en-US/firefox/addon/ublock-origin>): General-purpose ad blocker with custom rules set by the user.

4.3 Countermeasures Against Online Tracking Techniques

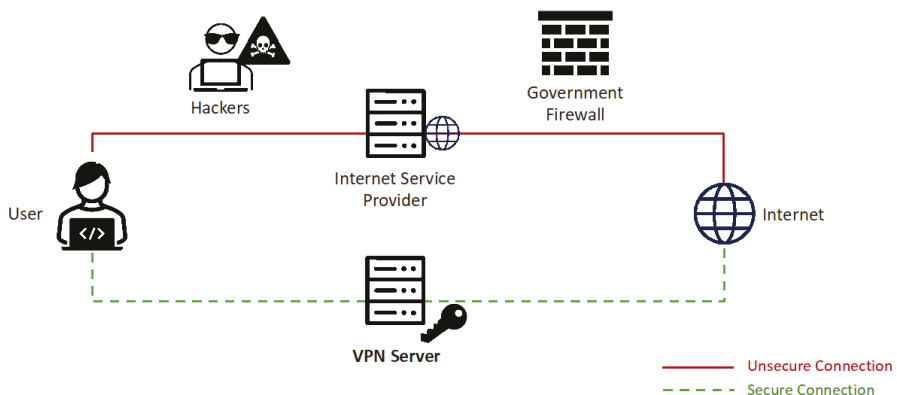
In the previous chapter, we discussed how online trackers employ various methods to track internet users' browsing activities. In this section, we present countermeasures for privacy invasion.

4.3.1 Conceal your IP address

Internet users can conceal their IP address using two methods:

- Using a virtual private network (VPN) (see Figure 4.4): A VPN works by establishing a secure encrypted tunnel between the user device and the VPN service. The VPN server will give the users a new IP address, making them appear connected from other geographical locations when surfing the internet. VPNs are used widely in the corporate world to secure connections to enterprise intranets when connecting over unsecured networks such as the internet.

Figure 4.4: How a VPN works.



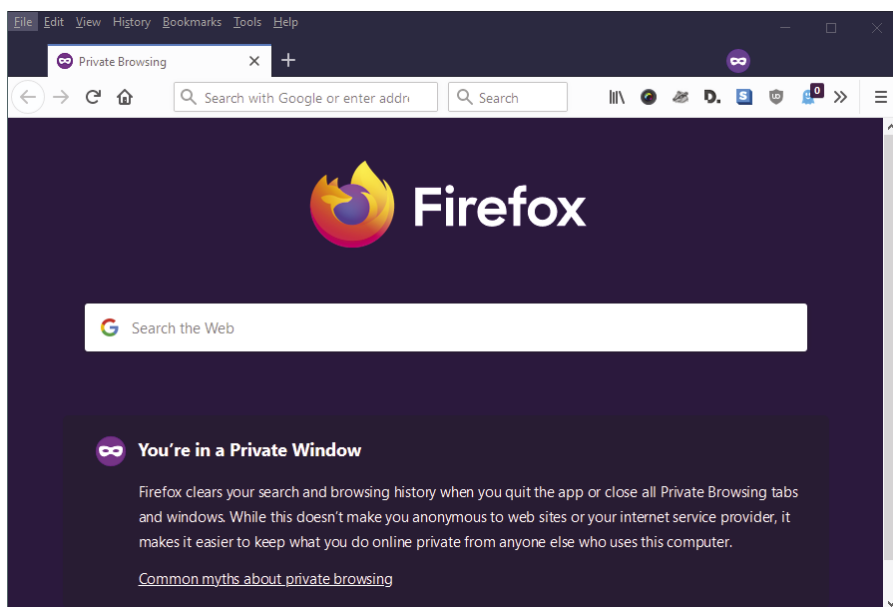
- An anonymity network, such as the Tor browser, can be used to surf the ordinary internet anonymously. This method is more secure than a VPN service and can guarantee a high level of online anonymity. Accessing the darknet will be covered in a dedicated chapter.

4.3.2 Cookie tracking

As we discussed in the previous chapter, cookies (especially first-party cookies created by the visited website) are helpful in facilitating many useful functions for users when visiting websites, such as remembering their login info and website theme settings. However, third-party cookies are the ones that raise privacy concerns as they can track internet users across multiple websites. It is recommended to block third-party cookies and delete web browser cookies when closing it.

All major web browsers, like Firefox and Chrome, have an incognito mode feature. When activated, the web browser will not record user browsing history. To activate this mode under Firefox, open Firefox and click CTRL + SHIFT + P. A new browser window will appear, stating you are in private browsing mode, as shown in Figure 4.5.

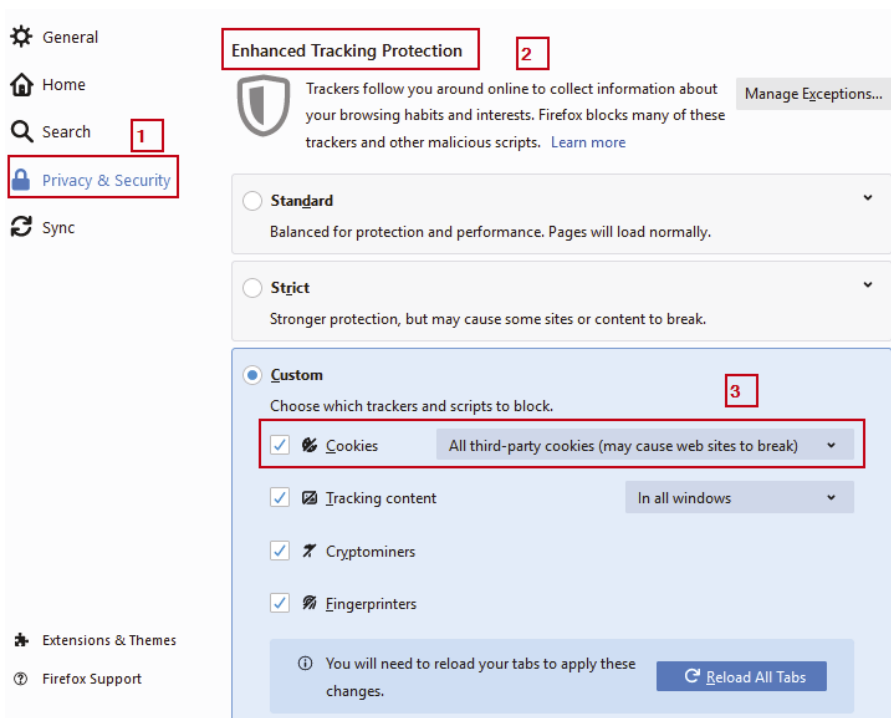
Figure 4.5: Firefox private mode browsing.



To stop third-party cookies under Firefox, follow these steps:

1. Go to the *Tools* menu >> *Options*
2. Select the *Privacy & Security* panel
3. Under *Enhanced Tracking Protection*, select the *Custom* radio button to choose what to block
4. To disable all third-party cookies, select the *Cookies* checkbox and select *All third-party cookies* (may cause web sites to break) from the drop-down, as shown in Figure 4.6.

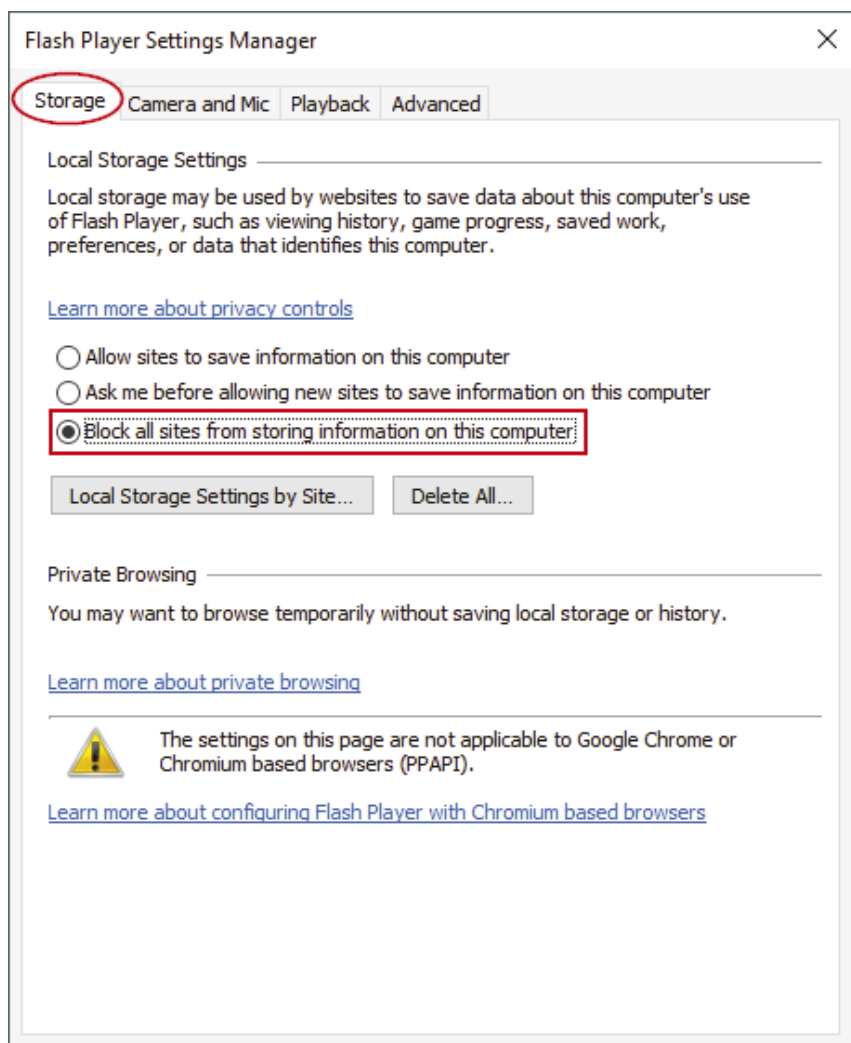
Figure 4.6: Disable third-party cookies under Firefox.



Flash cookies need a special arrangement to prevent them, as they are stored on the user's hard drive. To block Flash cookies under Windows, do the following:

Go to *Control Panel* >> *Flash Player* and select the option "Block all sites from storing information on this computer" (see Figure 4.7).

Figure 4.7: Block Flash cookies.



4.3.3 ETag tracking

To get rid of ETags, explained in Chapter 3, you must clear the browser cache content.

4.3.4 Digital fingerprinting

The only proven method to stop digital fingerprinting is to make your web browser similar to most internet users' web browsers fingerprints. By doing this, online trackers cannot uniquely distinguish your web browser fingerprint and thus cannot create a profile to record your web browsing activities. Of course, you still need to configure your web browser to block third-party cookies. You must also use a reliable VPN service to conceal your IP address.

To make your browser fingerprint similar to others, you should use a freshly installed web browser without installing any add-ons and run it inside a virtual machine, e.g., Virtual Box <https://www.virtualbox.org>. Using a virtual machine to run your browser inside it will also help you conceal browser fingerprint.

4.3.5 Use browser isolation technology

A new technology that is gaining more attraction recently among OSINT gatherers is browser isolation technology. In a nutshell, browser isolation is when a user uses a remote web browser hosted somewhere on the cloud to access web content.

Browser isolation provides a secure and controlled environment for conducting online research. This technology separates web browsing activities from the user's computing device, which significantly minimizes the risk of malware infections, data breaches, and exposure to malicious content. Browser isolation will also allow the user to hide their digital identity by concealing their IP address and prevent tracking their internet usage via digital fingerprinting.

4.4 Chapter Summary

The internet is a hostile environment where different observers are interested in tracking internet user's online activities. In this chapter, we gave technical advice on stopping this invasion and preventing online trackers from profiling your device when going online. Knowing how to conceal your online activities is critical for OSINT gatherers, as revealing their search activities may threaten to inform the entities they are researching.

Further Reading

1. Petsymposium, "Privacy Concerns and Acceptance Factors of OSINT for Cybersecurity: A Representative Survey" <https://petsymposium.org/popets/2023/popets-2023-0028.pdf> Accessed 2024-05-08
2. Techsafety, "Online Privacy & Safety" <https://www.techsafety.org/onlineprivacyandsafetytips> Accessed 2024-05-08
3. Privacytools, "Privacy guides" <https://www.privacytools.io/guides> Accessed 2024-05-08
4. Pcmag, "How to Completely Disappear From the Internet" <https://www.pcmag.com/how-to/how-to-stay-anonymous-online> Accessed 2024-05-08

CHAPTER

5

Open Source Intelligence (OSINT): A Practical Example

OSINT is the practice of gathering intelligence from publicly available sources to support intelligence needs. In the cybersecurity arena, OSINT is used widely to discover vulnerabilities in IT systems and is commonly named technical footprinting. Footprinting is the first task conducted by hackers – both black and white hat hackers – before attacking computer systems. Gathering technical information about the target computer network is the first phase in any penetration testing methodology.

In this chapter we will demonstrate how various OSINT techniques can be exploited to gain useful intelligence from public sources about a target IT environment.

5.1 Technical Investigation of a Website

The first step is to investigate a website. By knowing the type of programming language, web frameworks, and content management system (CMS) used to create the website, we can search for vulnerabilities that could target these components (especially zero-day vulnerabilities) and then work to exploit any of these vulnerabilities instantly once discovered.

There are different online services to examine the type of technology used to build websites. To use such a service, all you need to do is to provide the target domain name and you will get a full list of technical specifications and online libraries/programming language used to build a subject website. These

services also reveal the hosting provider of the target website, SSL certificate register name, and email system type. The following are some popular services:

- 1. BuiltWith (<https://builtwith.com>)
- 2. Wappalyzer (<https://www.wappalyzer.com>)
- 3. W3techs (<https://w3techs.com/sites>)
- 4. Genelify (<https://www.genelify.com/tools/technology-lookup>)

In the following screen capture, we use the *BuiltWith* service to investigate the technical specifications of a target website. This reveals different technical information, see Figure 5.1, and opens the door to more examination for each technology used to build the subject website. Now, we need to check the list of technical specifications to see whether there is unpatched operating system or outdated content management system with known vulnerabilities that can be exploited to gain access to the target system.

Figure 5.1: Using BuiltWith to investigate the technology used to build the target website.

Frameworks			
	ASP.NET 4.0	Nov 2013	Dec 2019
	ASP.NET	Jul 2012	Dec 2019
	Programming Language		
	ASP.NET MVC	Feb 2017	Dec 2019
	ASP.NET Ajax	Jul 2012	Feb 2017
			
Content Delivery Network			
	GStatic Google Static Content	Mar 2017	Nov 2019
Mobile			
	Viewport Meta	Mar 2017	Dec 2019
	iPhone / Mobile Compatible	Jun 2018	Dec 2019
Mapping			
	Google Maps for Work	Dec 2017	Dec 2019
			\$
	Google Maps API	Dec 2017	Dec 2019
	Google Maps	Dec 2017	Dec 2019
	Maps		
JavaScript Libraries and Functions			
	jQuery	Jul 2012	Dec 2019
	JavaScript Library		
	Modernizr 2.6	Feb 2017	Dec 2019
	jQuery UI	Mar 2017	Dec 2019
	jQuery Plugin - UI		
	Modernizr	Mar 2017	Dec 2019
	Compatibility		

For example, large numbers of ASP.net websites use *Telerik Controls* (<https://www.telerik.com>) to enrich their design. To find security vulnerabilities associated with *Telerik Controls*, you can go to <https://www.cvedetails.com> and search for *Telerik* security vulnerabilities (see Figure 5.2).

Figure 5.2: List of security vulnerabilities for Telerik controls.

Telerik : Security Vulnerabilities

CVSS Scores Greater Than: 0 1 2 3 4 5 6 7 8 9
Sort Results By : CVE Number Descending CVE Number Ascending CVSS Score Descending Number Of Exploits Descending
[View Results](#) [Download Results](#)

#	CVE ID	CWE ID	# of Exploits	Vulnerability Type(s)	Publish Date	Update Date	Score	Gained Access Level	Access	Complexity	Authentication	Conf.	Integ.	Avail.
1	CVE-2018-17060	22		Dir. Trav.	2018-10-08	2019-10-02	5.0	None	Remote	Low	Not required	Partial	None	None
Telerik Extensions for ASP.NET MVC (all versions) does not whitelist requests, which can allow a remote attacker to access files inside the server's web directory. NOTE: this product has been obsolete since June 2013.														
2	CVE-2017-11357	20		Exec Code	2017-08-23	2018-01-27	7.5	None	Remote	Low	Not required	Partial	Partial	Partial
Progress Telerik UI for ASP.NET AJAX before R2 2017 SP2 does not properly restrict user input to RadAsyncUpload, which allows remote attackers to perform arbitrary file uploads or execute arbitrary code.														
3	CVE-2017-11317	326		Exec Code	2017-08-23	2018-10-17	7.5	None	Remote	Low	Not required	Partial	Partial	Partial
Telerik.Web.UI in Progress Telerik UI for ASP.NET AJAX before R1 2017 and R2 before R2 2017 SP2 uses weak RadAsyncUpload encryption, which allows remote attackers to perform arbitrary file uploads or execute arbitrary code.														
4	CVE-2017-9248	522		XSS	2017-07-03	2019-10-02	7.5	None	Remote	Low	Not required	Partial	Partial	Partial
Telerik.Web.UI.dll in Progress Telerik UI for ASP.NET AJAX before R2 2017 SP1 and Sitefinity before 10.0.6412.0 does not properly protect Telerik.Web.UI.DialogParametersEncryptionKey or the MachineKey, which makes it easier for remote attackers to defeat cryptographic protection mechanisms, leading to a MachineKey leak, arbitrary file uploads or downloads, XSS, or ASP.NET ViewState compromise.														
5	CVE-2017-9140	79		XSS	2017-05-22	2018-09-27	4.3	None	Remote	Medium	Not required	None	Partial	None
Cross-site scripting (XSS) vulnerability in Telerik.ReportViewer.WebForms.dll in Telerik Reporting for ASP.NET WebForms Report Viewer control before R1 2017 SP2 (11.0.17.406) allows remote attackers to inject arbitrary web script or HTML via the bgColor parameter to Telerik.ReportViewer.axd.														
6	CVE-2015-2264			+Priv	2015-03-12	2015-03-13	6.9	None	Local	Medium	Not required	Complete	Complete	Complete
Multiple untrusted search path vulnerabilities in (1) EQATEC.Analytics.Monitor.Win32_vc100.dll and (2) EQATEC.Analytics.Monitor.Win32_vc100-x64.dll in Telerik Analytics Monitor Library before 3.2.125 allow local users to gain privileges via a Trojan horse (a) csunapi.dll, (b) swift.dll, (c) nfhwrchk.dll, or (d) surewarehook.dll file in an unspecified directory.														

There are many websites that list security vulnerabilities of operating systems, software and other web applications. The following are the most popular ones that we can use to search for common security vulnerabilities and exposures:

1.

<https://vulmon.com>

2.

<https://sploitius.com>

3.

<https://www.sauces.com>

4.

<https://www.shodan.io>

5.2 Analytics and Tracking

Most websites use Google services to analyze traffic and serve advertisements, we can use this feature to know all linked domain names. For example, we can find all websites that use the same *Google AdSense* or *Analytical* accounts by doing a reverse Google ID search. *Dnslytics* (<https://dnslytics.com/reverse-analytics>) is a free online service that finds domains sharing the same Google Analytics ID (see Figure 5.3).

Figure 5.3: Using the reverse Google Analytics service to reveal domain names belonging to the same entity.

The screenshot shows a web browser at <https://dnslytics.com/reverse-analytics>. The DNSlytics logo is at the top left, and a navigation bar includes 'Reports', 'Addons', 'Monitoring', 'Domain Tools', 'Reverse Tools', and 'More'. Below the navigation bar, a breadcrumb trail reads 'Home -> Reverse Tools -> Reverse Analytics -> powerofdiscussion.com'. The main content area is titled 'Reverse Google Analytics' and contains the instruction 'Find domains sharing the same Analytics ID.'. There is a text input field with a blurred domain name and a 'Go' button. Below the input field, a hint text says 'Enter domain name or Analytics ID. Example: qrutls.com or ua-15589237'. Below this, a light blue box displays 'Found Analytics ID: ua-15589237 for powerofdiscussion.com'. At the bottom, another light blue box states 'Found 5 domains using Analytics ID: ua-15589237', with the number '5' circled in red.

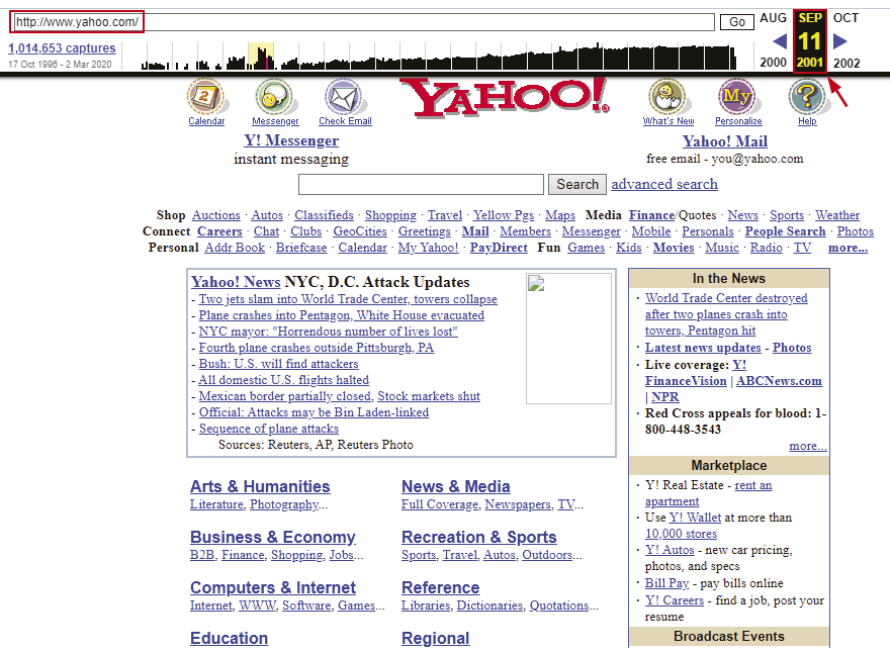
In the same way, we can use source code search engines to find Google service IDs and other code snippets across the web. Here are some free code search engines:

1. Codase (<https://codase.com>)
2. Search code (<https://searchcode.com>)
3. Nerdydata (<https://www.nerdydata.com/reports/new>)

5.3 Website History

In many instances, checking the old version of the website can reveal important information. For example, an old website version of a corporate may reveal top management email addresses and phone numbers before they got removed from the new version. *Wayback Machine* (<https://archive.org/web>) is a good place to start your search for old versions of websites (see Figure 5.4).

Figure 5.4: Using the Wayback Machine to see previous versions of websites.



Other online services for finding previous versions of websites are:

1. Archive.today (<https://archive.ph>)
2. ArchiveBox (<https://archivebox.io>)
3. Memento Time Travel – Chrome extension (<https://mementoweb.org/about>)
4. PublicWWW (<https://publicwww.com>)

5.4 Subdomain Discovery

Revealing a website's subdomain is critical in the OSINT gathering process. Subdomains can reveal a significant amount of sensitive information about the target organization's IT environment, enabling threat actors to better understand the target's IT infrastructure and potential attack surfaces.

Discovering subdomains can reveal various IT systems and services that may not be directly accessible or visible from the organization's primary domain.

For instance, some examples of sensitive information that can be discovered through subdomain enumeration include:

1. VPN portal: Organizations commonly use separate subdomains for their virtual private network (VPN) portals. As we know, VPN facilitates remote access to internal resources. Discovering the VPN subdomain can provide an entry point to execute more cyber-attacks by exploiting vulnerabilities in the VPN solution.
2. Email systems: Subdomains pointing to email services, such as mail.example.com or web-mail.example.com, can reveal the organization's email systems (e.g., whether it is open source or commercial). This information is valuable for executing phishing attacks.
3. File transfer protocol (FTP) server: FTP servers are often hosted on separate subdomains (ftp.example.com) and may contain sensitive files or directories inadvertently left unprotected with a password.
4. Development environments: Many organizations use dedicated subdomains for their development and testing environments when developing in-house software solutions (dev.example.com, staging.example.com). These sub-domains may contain unpatched vulnerabilities or sensitive information about the target company's IT infrastructure and software solutions that attackers can exploit.
5. Content management systems (CMS): The subdomains of a company can reveal its usage for a particular CMS, such as Wix or WordPress. Attackers can exploit this knowledge to find vulnerabilities in the installed version of the CMS.

To find all subdomains of a target indexed by Google, use the Google search command shown in Figure 5.5.

Figure 5.5: Replace example.com with your target domain name.



We can also use the following Google search queries to find all subdomains associated with a website:

site:example.com inurl:subdomain | This query identifies subdomains based on URL patterns.

site:example.com intitle:subdomain | This query reveals subdomains appeared in page titles

site:example.com -site:www.example.com -site:example.com | This query identify subdomains other than www.

site:example.com filetype:xml | This query searches for all XML files hosted on a particular domain name. XML files may contain references to hidden subdomains.

5.5 Type and Versions of IT Infrastructure of the Target Company

Job websites and any job announcement posted on the target website should be analyzed to discover the exact IT infrastructure used by the target organization. For example, we conducted a simple search on employee resumes on job websites. We were able to capture important information about given organization's security systems (e.g., firewalls and intrusion detection systems), server operating system types, email systems, networking devices, types of backup systems and much more (see Figure 5.6).

Figure 5.6: Sample resume found on a job website that reveals the type of IT infrastructure of the target organization.

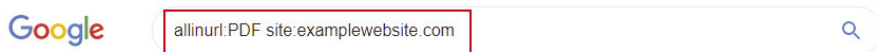
- Manage AD services on Windows Server 2008, 2012, 2016 (DNS and DHCP)
 - Configure user environment and Implement Security by using GPO.
 - Manage Microsoft Exchange 2010, 2013 and 2016 (Mailbox, Mailbox Policy, OWA, And SSL Certificate).
 - Backup Critical servers (Exchange, SQL, Oracle, DC and System State) using Veritas Backup Exec.
 - Manage Microsoft Office 365 and Yammer.
 - Configure and manage virtual machines on Hyper-V
 - Manage Security appliance: Barracuda, Fortinet and SonicWall
 - Manage Servers and clients, Deploy OS and install application using (SCCM 2012).
 - Monitor Healthy and performance of servers and application using (SCOM 2012).
 - Manage VMs, Monitor performance and resources of Hyper-V hosts using (VMM 2012).
 - Setup and Manage Enterprise Antivirus (Symantec, Kaspersky)
-

5.6 Harvest Digital Files Hosted on Domains

Using advanced Google search engine techniques (also known as Google dorks) can reveal a significant amount of information about an organization's IT systems in addition to confidential files left on the public server. There are thousands of Google dorks, and you can practice creating your own. A comprehensive list of Google dorks can be found in the Google hacking database (<https://www.exploit-db.com/google-hacking-database>).

We will experiment using Google dork to locate all PDF files posted on the target website (see Figure 5.7):

Figure 5.7: Find all PDF files on the target domain name.



In the above example, we searched for PDF files; however, you can change the file type, e.g., doc, docx, xls, txt, etc.

Here are more Google dorks to find different types of files hosted on a particular website:

site:example.com filetype:doc OR site:example.com filetype:docx | Find MS Word documents

site:example.com filetype:xls OR site:example.com filetype:xlsx | Find MS Excel documents

site:example.com filetype:ppt OR site:example.com filetype:pptx | Find MS Powerpoints documents

site:example.com filetype:bak OR site:example.com filetype:old | Find backups files

site:example.com filetype:php OR site:example.com filetype:js OR

site:example.com filetype:java OR site:example.com filetype:c OR

site:example.com filetype:cpp | Find source code files

site:example.com

filetype:pdf|doc|docx|xls|xlsx|ppt|pptx|txt|zip|rar|sql|db|xml|config|ini|bak|old|log|php|js|java|c|cpp This is a single query that combines all previous search queries to find all popular file types hosted in a particular domain name

5.7 Information Contained in File Metadata

We should investigate the metadata of each file found on the target website. Metadata can be defined as data about data, in technical terms metadata contains hidden descriptive information about the file it belongs to. For example, some metadata in an MS Office document file might include the author's name, date/time created, comments, software used to create the file, and the type of OS of the device used to create the file. (see Figure 5.8).

Figure 5.8: Checking PDF file metadata info.

Document Properties

Description	Security	Fonts	Initial View	Custom	Advanced
Description					
File:	elearning_user_manual.pdf				
Title:	PowerPoint Presentation				
Author:	Perry Longman				
Subject:					
Keywords:					
Created:	7/3/2017 5:16:05 PM				
Modified:	7/3/2017 5:16:05 PM				
Application:	Microsoft® PowerPoint® 2010				
Advanced					
PDF Producer:	Microsoft® PowerPoint® 2010				
PDF Version:	1.5 (Acrobat 6.x)				

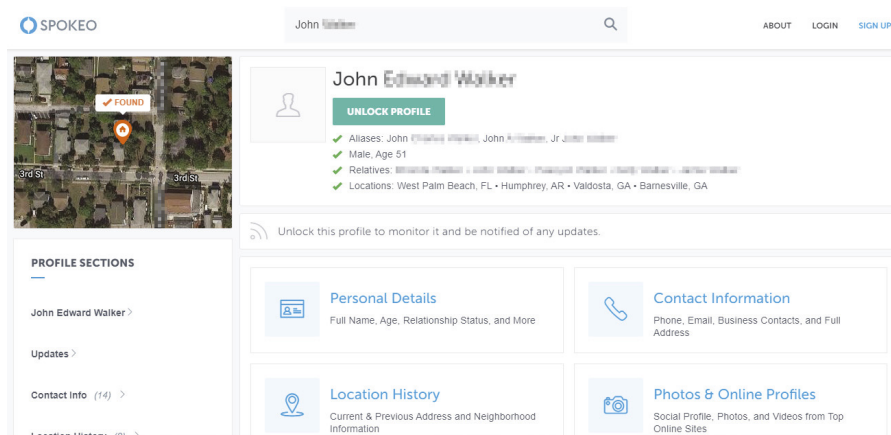
From Figure 5.8, we found the following facts about the subject PDF file metadata:

1. Installed PDF reader Version on the creation device: **1.5**
2. Application used to create the report: **MS PowerPoint 2010** (using the "Save As" function)
3. Type of OS used on the target device: **Windows**
4. File creation date/time: **July 2017**
5. **Author Name** (the person who creates the file).

If the file contains the author's name, using specialized people data collection websites, we can expand our search to look up more details of the file's author. The following are some of the well-known people's search engines:

1. Spokeo (<https://www.spokeo.com>) (see Figure 5.9)
2. Truepeoplesearch (<https://www.truepeoplesearch.com>)
3. Truthfinder (<https://www.truthfinder.com>)
4. 411 (<https://www.411.com>)

Figure 5.9: Using SPOKEO to look up information about people you know.



5.8 Tools to Retrieve Digital File Metadata

Digital files are not limited to PDF and MS Office documents; during OSINT investigations, we may encounter images and video files. There are numerous tools to inspect digital files metadata. Here are three popular free tools:

- ExifTool (<https://exiftool.org>) – inspect images metadata (EXIF metadata)
- Exif Pilot (<https://www.colorpilot.com/exif.html>)
- AnyMP4 Video Converter (<https://www.anymp4.com/video-converter-ultimate>) Video metadata editor

5.8.1 Email naming criteria

To predict the naming criteria a given organization uses when creating email accounts, we should investigate the naming criteria used for current email addresses. For example, many organizations use the following naming criteria:

- Most common patterns of naming new emails: **{first}{DOT}{last first three characters}@exampleWebsite.com**
- Other naming criteria include: **{first}@exampleWebsite.com**.

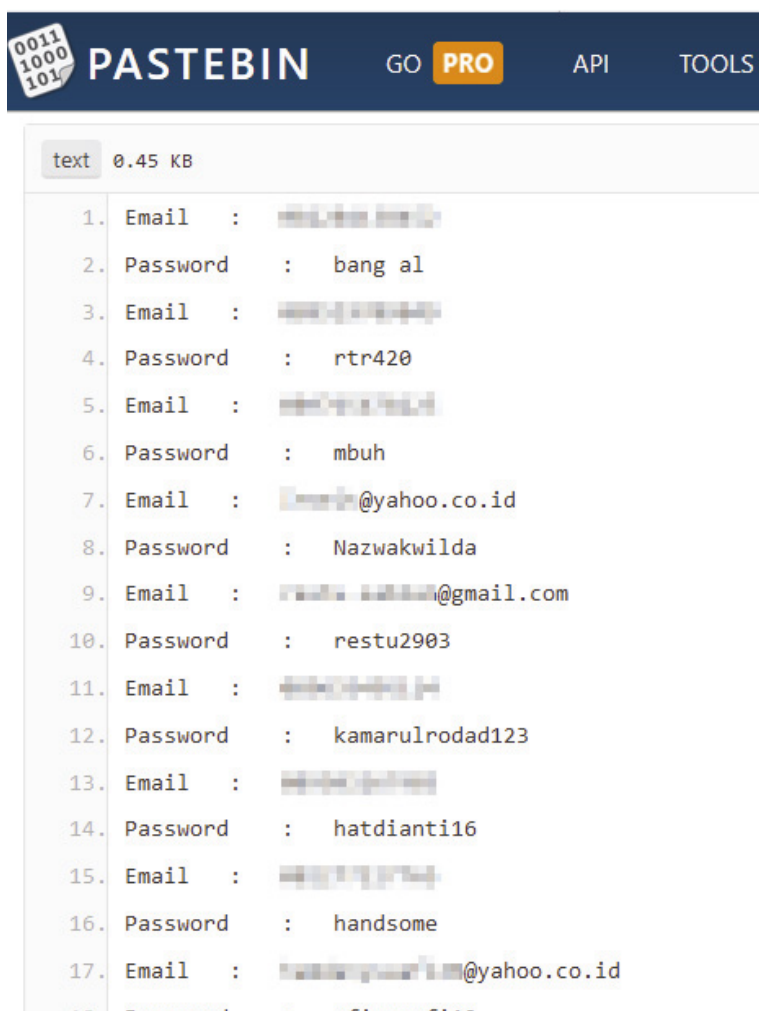
We usually use <https://www.email-format.com> to find the email address formats used by several companies. Below are additional email investigation tools:

- Hunter.io (<https://hunter.io/email-verifier>) – email verification tool
- theHarvester (<https://github.com/laramies/theHarvester>) – a free command line tool for finding a particular email address across the internet
- Whoxy (<https://www.whoxy.com>) – find all domain names linked to an email address.

5.8.2 Leaked credentials

Leaked account credentials are spread everywhere online, especially in the darknet. For example, Pastebin websites (see Figure 5.10) contain a vast amount of leaked credentials. Have I Been Pwned (<https://haveibeenpwned.com>) lets you know if your email address was included in a previous data breach.

Figure 5.10: Leaked credentials found on Pastebin.com.



5.9 Chapter Summary

This chapter presented an overview of various OSINT capabilities and how to gather valuable intelligence about different entities. In today's information age, having OSINT skills is very valuable. However, there are many prerequisites you should master to make your OSINT search rich and compelling. For instance,

OSINT is strongly related to digital forensics, so knowing basics digital forensics operations will also prove helpful when conducting OSINT related information gathering activities. OSINT gatherers should also have a fair understanding of social media research techniques to develop a methodology for collecting intelligence from different social media platforms.

In the coming three chapters, we will discuss three critical topics in OSINT research which are:

1. Using machine learning (ML) and artificial intelligence (AI) in OSINT research
2. Developing a methodology to collect intelligence from social media platforms
3. Searching within the deep and darknet.

In the next chapter, we will continue the discussion of using different OSINT techniques to find intelligence from public sources; however, we will begin discussing how to leverage the latest technological advances in machine learning (ML) and artificial intelligence (AI) tools in our OSINT search process.

Further Reading

1. Bellingcat, "Automatically Discover Website Connections Through Tracking Codes" <https://www.bellingcat.com/resources/2017/07/31/automatically-discover-website-connections-tracking-codes> Accessed 2024-05-08
2. Gralhix, "List of OSINT Exercises – Challenge Yourself!" <https://gralhix.com/list-of-osint-exercises> Accessed 2024-05-08
3. Secjuice, "A Guide To Social Media Intelligence Gathering (SOCMINT)" <https://www.secjuice.com/social-media-intelligence-socmint> Accessed 2024-05-08



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER

6

Using AI in OSINT Research

After the public release of ChatGPT, the usage of generative artificial intelligence (AI) technology has witnessed huge attention. Generative AI tools provide numerous advantages for businesses and individuals alike that range from summarizing large amounts of text to writing programming code and generating documentation for software projects with ease. However, for OSINT researchers, AI can play a critical role in improving many aspects of their work.

AI technology can enhance various aspects of OSINT gathering, from data collection and analysis to report generation and decision support. By leveraging AI capabilities, OSINT researchers can process and derive insights from large volumes of data more efficiently, identify relevant patterns and relationships that may be difficult to detect manually, and finally make more informed decisions in their investigations, which can ultimately lead to a better intelligence product.

In this chapter, we discuss a few use cases on how OSINT researchers can exploit various AI tools in their research to streamline work processes and enhance gathering capabilities.

6.1 Data Collection and Scraping

The first task of an OSINT gathering activity is data collection. OSINT researchers may need to collect considerable volume of data about their targets. This data is collected from various online sources, such as:

- Social media platforms – such as Facebook, Instagram, Twitter (now X) and YouTube
- Job posting websites such as Indeed and LinkedIn

- News websites, including traditional media websites (magazines, radio stations, and TV websites)
- Discussion forums – such as Reddit
- Government databases – such as viral databases, business filings, and property/criminal records
- Pastebin and any website housing breached content
- Whois websites to reveal websites and IP address ownership.

Collecting data from all these sources would be daunting for OSINT researchers. Traditional web scraping tools require pre-defined selectors to be set up first that cannot be adjusted after the web scraping activity begins. By leveraging AI tools, they can automate the web scraping process to a great extent, in addition to customizing their gathering activities according to each search case. For instance, AI algorithms are capable of adjusting their work during scraping to gather data from dynamic websites without human intervention.

For example, if an investigation target was a drug dealer group, we could configure the AI web scraping tool to look for all data related to this group across the web, including forums, social media, news reports, and dark web marketplaces. This targeted search will lower the amount of data we need to gather, in addition to returning more focused results. The AI web scraper can also continually monitor the web for any mention related to the searched entity and send notifications in real-time when detecting any change or mention.

Here are links to some AI web scraping tools:

- AnyPicker (<https://app.anypicker.com>) – this is a free web data scraper Chrome extension
- Browse AI (<https://www.browse.ai>) – extract and monitor data from any website
- Diffbot (<https://www.diffbot.com>) – offers AI-powered web scraping and data extraction services.

6.2 Analysis of Unstructured Text Data

A significant obstacle faced by OSINT researchers is collecting unstructured data. Unstructured data refers to all data that does not conform to a pre-defined data model or is not contained in a specific organized way – such as within a database management system. Examples of unstructured data include:

- Social media website posts
- Legal and business documents

- Presentations
- Images, videos, and audio content existing on websites
- Emails.

Natural language processing (NLP) is a subset of AI technology that allows computer systems to understand, generate, and process human language. Gaining insight into unstructured data is an important aspect of many OSINT investigations.

NLP can benefit OSINT gathering in multiple ways:

- Speech recognition
- Summarizing text
- Text classification
- Sentiment analysis
- Topic modeling
- Named entity recognition.

NLP tools and techniques can be exploited to gain insight from unstructured data to support our intelligence needs. For instance, text analytics can be used to extract keywords from a large volume of textual data to identify named entities such as a company, a hacker group, or an individual.

Sentiment analysis is another case of AI being used to support OSINT investigations. Using sentimental analysis tools, OSINT gatherers can easily identify the emotional tone of social media conversations such as happiness, fear, or anger. The global market of sentimental analysis tools is predicted to reach⁴ US\$4.84 billion by 2026. Here are some open source sentimental analysis tools that OSINT researchers can use to analyze text content:

- spaCy (<https://spacy.io>) supports more than 75+ languages and can extended with custom components and attributes.
- NLP.JS (<https://github.com/axa-group/nlp.js>) supports 40 languages, provides sentiment analysis for phrases and named entity recognition and management.
- Pattern (<https://github.com/clips/pattern>) provides natural language processing in addition to data mining for the web services Google, Twitter, and Wikipedia.

⁴Mordorintelligence, “Text Analytics Market Size & Share Analysis-Growth Trends & Forecasts (2024–2029)”, <https://www.mordorintelligence.com/industry-reports/text-analytics-market>” Accessed 2024-05-01.

6.3 Analysis of Multimedia Files (Images and Videos)

AI-powered tools can collect and analyze images/video files published online on a massive scale to extract named entities – such as a person, a company, or a group of people. Later, we can use specifically trained ML models to aid in image/video processing to achieve the following results:

- First, AI tools can facilitate collecting media files from various online sources, including social media platforms, websites, and specialized databases.
- Classify and categorize images automatically into groups or segments based on content, subjects, or other relevant criteria.
- Perform facial recognition on images – for example, by using deep learning technologies such as convolutional neural networks (CNNs)⁵ to identify facial features like eyes and noses.
- Divide the collected image into separate regions and analyze each region individually, allowing for a more detailed examination of specific areas of interest.
- Enhance the quality of collected images – for example, by using AI tools to upscale or magnify images to improve their clarity and detail. Upscale.media is a powered AI tool to enhance image quality.
- Identify fake images or those created using deepfake technology, which can be crucial for verifying the authenticity and integrity of visual evidence.
- Search across the web and specialized databases for similar images to a specific one, aiding in identifying potential connections or related content.
- Analyze video files by extracting keyframes, detecting objects and activities, generating transcripts from audio, and identifying individuals through facial recognition.

The field of using AI and ML in multimedia analysis is evolving rapidly. We expect to see more techniques and tools to aid OSINT researchers soon.

6.4 Content Summarization

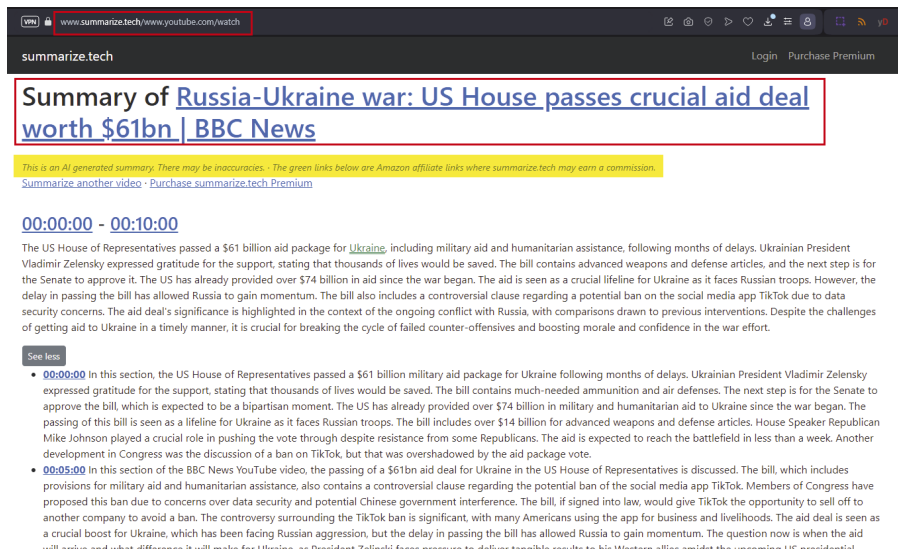
A regular task of OSINT gatherers is summarizing large chunks of text. For instance, we may encounter a large number of business documents that need to be summarized or extract specific keywords from them – such as names, emails, and phone numbers.

⁵Researchsquare, “Convolutional Neural Networks for Face Recognition: A Systematic Literature Review” https://assets.researchsquare.com/files/rs-3145839/v1_covered_39e340dd-f852-4351-86c4-1b530d637532.pdf Accessed 2024-04-26

Content summarization is not only limited to text content; for example, AI-powered tools can be used to summarize videos. Here are some links to AI-powered content summarization tools:

- Linkquire (<https://www.linkquire.com>) – summarize YouTube videos. You can also ask the tool questions about the video content, and AI will provide the answers directly from the video.
- Recall (<https://app.getrecall.ai>) – summarize different types of online content.
- SummarizePaper (<https://summarizepaper.com>) – summarizes scientific papers from the arXiv platform.
- Summarize Tech (www.summarize.tech) – summarize YouTube videos (see Figure 6.1).

Figure 6.1: Using summarize.tech AI-powered tool to summarize YouTube videos.



6.5 Social Network Analysis

Social media platforms are generating massive volumes of content every day. Most OSINT investigations require collecting data from these platforms. AI can help in this aspect very well by discovering relationships, revealing connections, and finding influential persons within online communities. For instance, AI-powered tools can aid in social media investigations via the following areas:

- Identify key persons within online communities: AI-powered tools can quickly identify prominent persons within online communities (by inspecting their number of followers and/or the number of replies and shares in their posts and comments).
- Reveal disinformation campaigns: Disinformation or spreading fake information has become a top concern for governments, companies, and individuals. AI-powered tools can reveal social media accounts responsible for spreading such fake content and how they are connected with other accounts. This facilitates revealing the origin sources of misinformation or propaganda.
- Group social media users: AI tools can help group individual users based on their connections, interests, ideologies, and participation in social media communities. This allows OSINT researchers to understand those social groups' landscape and potential risks.
- Topic modeling: AI can help identify the main topics discussed within each community. This allows OSINT gatherers to discover trending topics and understand what each group or community cares about. On the other hand, AI-powered tools can help in finding online communities where extremist or terrorist organizations are trying to recruit new persons.
- Data verifications: Checking the validity and authenticity of the information is a vital skill for OSINT gatherers. AI-powered tools can help verify and cross-reference gathered information to ensure its validity for the investigation. For example, an AI system can analyze the metadata of images and videos to validate source credibility and content consistency to assess the reliability of gathered information. This allows OSINT analysts to separate factual information from potential misinformation spread by malicious parties.

6.6 Chapter Summary

The use cases presented in this chapter demonstrate how AI and ML systems can enhance various aspects of OSINT gathering, from data gathering and analysis to reaching decision support. By leveraging AI capabilities, OSINT researchers can process and derive deep insights from large volumes of data (both structured and unstructured) more efficiently, identify common patterns and relationships that may be difficult to detect manually, and finally make more informed decisions in their investigations.

Further Reading

1. Theverge, "OpenAI working on new AI image detection tools" <https://www.theverge.com/2024/5/7/24151482/openai-image-detection-ai-watermarking-audio> Accessed 2024-05-08
2. Arxiv, "Detecting AI-Generated Images via CLIP" <https://arxiv.org/abs/2404.08788> Accessed 2024-05-08
3. Makeuseof, "The 8 Best AI Image Detector Tools" <https://www.makeuseof.com/ai-image-detector-tools> Accessed 2024-05-08
4. authentic8, "Using OSINT to identify AI-generated content" <https://www.authentic8.com/blog/osint-ai-generated-content> Accessed 2024-05-08

CHAPTER

7

Social Media Intelligence (SOCMINT)

Social media intelligence (SOCMINT) is a sub-branch of OSINT. It includes all tools, techniques, and methods used to gather and analyze information from social media platforms. It is worth noting that SOCMINT focus is on investigating social media platforms as a whole, hence, not only social networking websites such as Facebook and Instagram. For instance, when executing SOCMINT, OSINT researchers need to consider all social media platforms, such as video-sharing sites like YouTube, image-sharing websites like Pinterest and Flickr, microblogging sites such as Twitter (X platform) and Mastodon in addition to social sites that resemble traditional discussion forums such as Reddit.

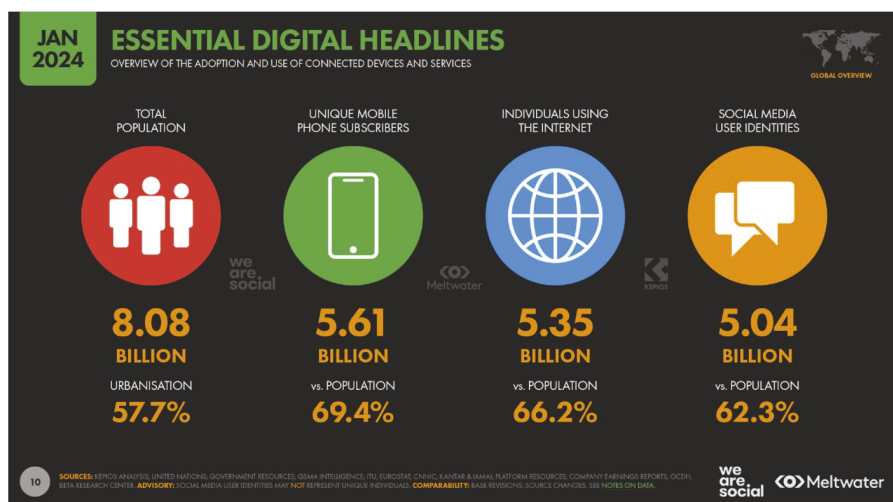
The number of social media users worldwide is increasing at an explosive rate. According to Datareportal⁶, there are 5.35 billion internet users worldwide, and 5.03 billion of them are social media users. The global number of social media users is expected to increase in the coming years to reach six billion in 2027⁷. This makes SOCMINT a vital area for OSINT investigators to inspect (see Figure 7.1).

The interactions on social media platforms take various forms. For example, between individual users (or user to user), between a user and a group or online community, or interactions between two groups.

⁶ Datareportal, “DIGITAL 2024: GLOBAL OVERVIEW REPORT”, <https://datareportal.com/reports/digital-2024-global-overview-report> \Accessed\2024-04-23

⁷ Statista, “Number of social media users worldwide from 2017 to 2027” <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users> \Accessed\2024-04-24

Figure 7.1: The global number of social media users worldwide. Source: <https://datareportal.com/reports/digital-2024-global-overview-report>



Social media users post different types of content to their online profiles, such as:

- Images
- Videos
- Text posts
- Links to other content – such as news articles or YouTube videos
- Location check-ins – such as the location of a restaurant where a user was eating lunch.

Digital files (such as images and videos) posted on some social media platforms can be inspected to reveal more hidden information. This info is called Metadata (e.g., author name, date/time when the file is created, type of capturing device), and some social media platforms, like Facebook and Twitter (X) strip metadata from digital files before uploading them to user profiles.

In this chapter, we discuss the concept of SOCMINT and suggest a general methodology to search within social media platforms. We will provide links to various tools and online services for executing advanced searches on popular social media platforms. However, before we begin, let us differentiate between OSINT and SOCMINT regarding privacy.

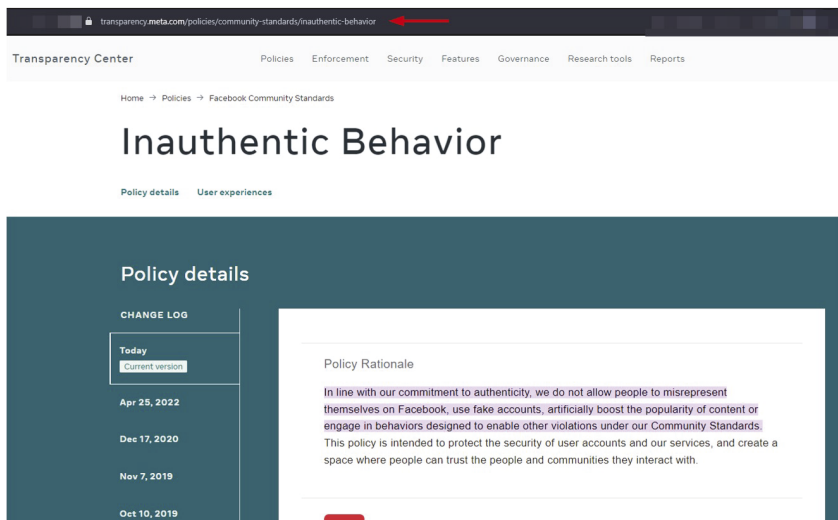
7.1 Privacy Issues In SOCMINT

As already discussed in Chapter 1, OSINT resources include all information available from public sources. Although some data needs to be purchased first, such as digital library subscriptions or access to some corporate filings, it is still considered open source data because it is accessible to the public.

Now, when it comes to SOCMINT, information available on social media platforms is also public. However, most social media users consider their profile content to be private and do not wish other entities to exploit them for marketing or intelligence purposes. This raises ethical concerns and legal implications regarding using personal data for intelligence purposes.

SOCMINT is more complex in terms of privacy because OSINT researchers may need to create fake profiles or use automated tools to scrape and collect data from social media platforms. Creating fake profiles is against most social media platforms' policies (see Figure 7.2). In addition to this, social media platforms often have strict terms of service and data usage policies that may prohibit or limit the automated collection and analysis of user data for commercial or intelligence purposes. OSINT gatherers must carefully navigate these

Figure 7.2: The Meta company prohibits the creation of fake accounts on all its platforms, such as Facebook and Instagram.



legal and ethical boundaries to ensure compliance and avoid potential legal consequences.

Regardless of internet users' privacy expectations, OSINT gatherers will undoubtedly exploit the information available on social media platforms because they cannot simply ignore it since it can provide a treasure trove of information for their investigations.

7.2 OSINT Roadmap for Investigating Social Media Platforms

Creating a detailed roadmap for investigating each social media platform requires a book on its own. So, for the remainder of this pocket guide we will propose a roadmap suitable for researching different social media platforms. However, we will introduce links to tools and online services to further investigate particular features of popular social media platforms.

There are four main phases for executing SOCMINT on any social media platform.

7.2.1 Phase 1: Preparation and setup

The initial step of any SOCMINT investigation begins with preparation.

Three steps of preparation

First, we need to identify the scope of our research and the entities involved in it. For example, are we going to investigate a company, an individual, or a group of people – such as a drug dealer group. Another example could be tracking a particular event. For example, police may leverage social media platforms to monitor a specific event before it occurs to detect any problems, such as disturbances or potential protests.

The second thing we need to identify in the first phase is the social media platforms we want to inspect. For example:

- To monitor a student gathering activity, we need to focus on social media platforms used by matures, such as Facebook, Instagram, and Twitter (X).
- If we are tracking a group of hackers, then focusing on social platforms that provide a level of anonymity (due to their privacy features or the difficulty in tracing user identities) is preferred, such as Reddit, Mastodon, and Telegram.

The third step in the first phase is to set up profiles on target social media platforms. As we already mentioned earlier, creating a fake profiles is against most social media websites' policies, so make sure to use these accounts wisely and document your work while using them.

There are different online services for creating fake social media accounts. Fake Name Generator (<https://www.fakenamegenerator.com>) is one of them.

Setup searching device

We need to set up the computing device used to execute our research. Preparing the computing device involves installing proper research tools and installing the required anonymity software such as VPN and the TOR browser to browse the darknet.

There are many out-of-box security operating systems explicitly configured for executing OSINT research. These are commonly Linux-based systems and come loaded with tons of OSINT and other security tools. Here are the main prominent ones:

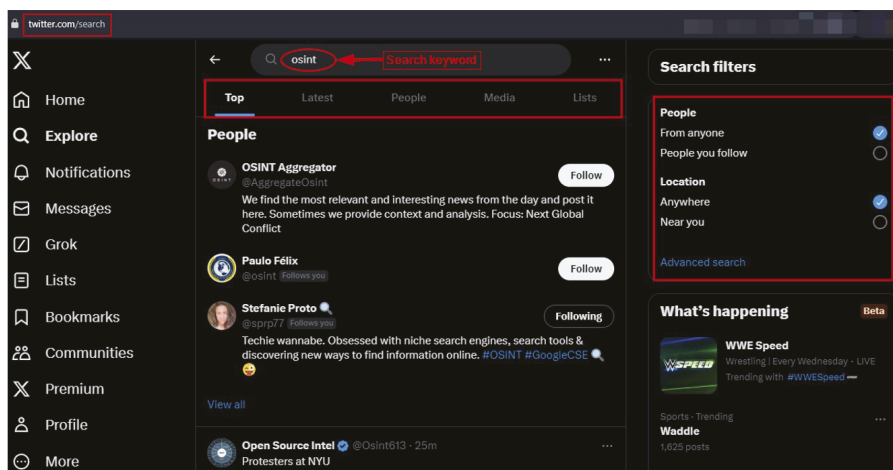
- Tsurugi Linux (https://tsurugi-linux.org/tsurugi_linux.php)
- Trace Labs OSINT VM (<https://www.tracelabs.org/initiatives/osint-vm>)
- CSI Linux (<https://csilinux.com>)
- Osintux (<https://www.osintux.org>)
- Kali Linux (<https://www.kali.org>) is heavily geared towards penetration testing and ethical hacking, but it also has a dedicated suite of OSINT tools for information gathering.

7.2.2 Phase 2: Data collection

This is the primary phase of SOCMINT, where OSINT researchers gather the raw data for analysis. The data collection phase involves the following sub-steps:

Performing manual search

Some OSINT investigation cases don't require a broad search scope. In such a case, we can use manual methods to search for keywords, geolocation data, and hashtags in the target social media platform. For instance, the Twitter (X) platform has an internal search functionality for finding information within the platform. Twitter search results can be filtered according to different criteria to narrow search results (see Figure 7.3).

Figure 7.3: Twitter's internal search functionality can be filtered according to different criteria.

If you want to return more specific results, you can utilize advanced Twitter search operators to refine your search queries. A detailed guide of Twitter operators and examples of how to use each one can be found in Twitter's official documentation "Rules and filtering: Standard v1.1" (<https://developer.twitter.com/en/docs/twitter-api/v1/rules-and-filtering/search-operators>).

Reddit is another popular social media website that uses the same structure of discussion forums. Reddit is composed of communities (also known as subreddits), and each subreddit focuses on a specific topic, such as technology, nature, sport, food, cooking, and education, to name only a few. You can find lists or a directory of subreddits posted on the Reddit website (<https://www.reddit.com/r/findareddit/wiki/directory>). However, some online services (such as Redditlist (<https://redditlist.com>)) list all subreddits along with some information about each one, such as the number of subscribers, rank, and recent activity (see Figure 7.4).

Social media monitoring tools

These tools monitor social media interactions for specific mentions, hashtags, comments, or keywords across different social media platforms. OSINT gatherers will find it daunting to manually track social media mentions across different websites. Likely, there are automated tools to facilitate this action.

Figure 7.4: redditlist.com lists all subreddits or communities that exist on the Reddit platform.

The screenshot shows the redditlist.com website with three main data tables. The top navigation bar includes the redditlist logo and links for /r/REDDITLIST and SUBREDDITS. Below the navigation bar are filters for All, SFW, and NSFW. The three tables are:

Recent Activity			Subscribers			Growth (24Hrs)		
Rank	Subreddit	Subscribers	Rank	Subreddit	Subscribers	Rank	Subreddit	Growth
1	AskReddit	23,296,387	1	announcements	41,411,622	1	WizardsUnit	19.39%
2	news	18,436,141	2	funny	25,058,374	2	crashteamracing	19.1%
3	funny	25,058,374	3	AskReddit	23,296,387	3	KpopSexy	17.67%
4	worldnews	21,536,704	4	gaming	22,647,964	4	harrypotterw	16.64%
5	politics	5,203,312	5	pics	22,175,710	5	EvilE	16.49%
6	aww	20,996,624	6	science	21,758,637	6	iamanutterpieceofshit	15.76%
7	pics	22,175,710	7	worldnews	21,536,704	7	CompetitiveTFT	15.27%
8	todayilearned	21,033,667	8	todayilearned	21,033,667	8	BestTeensNSFW	15.17%
9	gaming	22,647,964	9	aww	20,996,624	9	FullPorn	12.73%
10	videos	20,704,486	10	movies	20,916,873	10	whiptax	11.73%
11	nba	2,438,975	11	videos	20,704,486	11	underlords	11.45%
12	tfu	14,499,101	12	Music	20,350,355	12	nude_snapchat	10.46%
13	AmtheAsshole	929,154	13	AMA	19,146,065	13	EllieLouPics	9.45%

Social media monitoring is not limited to social mentions of your name or brand. For example, it can provide a good insight into your competitor's strategies and future plans in addition to discovering the sources of wealth for both individuals and corporations.

Here are four social monitoring tools:

- Hashtag Search (<https://postcron.com/en/blog/landings/hashtag-search-tool>)
- Brandwatch (<https://www.brandwatch.com>)
- Sprout Social (<https://sproutsocial.com>)
- Mention (<https://mention.com/en>)
- HubSpot (<https://www.hubspot.com>)
- Google Alerts (<https://www.google.com/alerts>).

7.2.3 Phase 3: Analysis

The third phase of our SOCMINT search plan analyzes the gathered data to extract meaningful insights and intelligence. OSINT researchers can utilize the following tools and techniques:

Use natural language processing (NLP) techniques

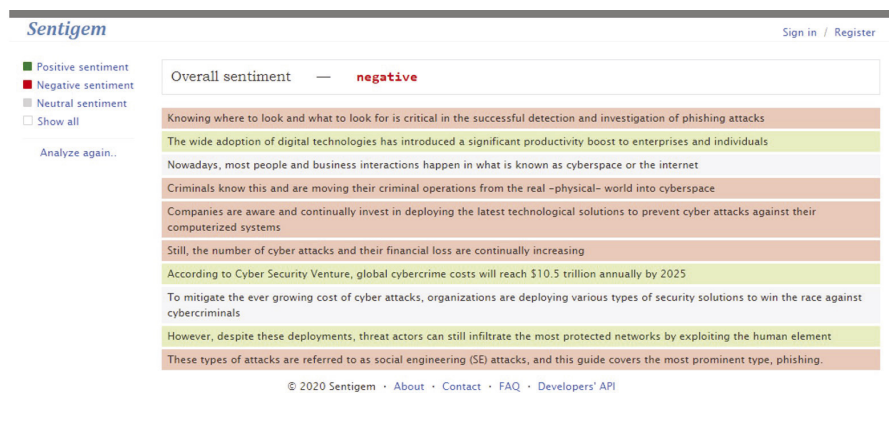
NLP facilitates analyzing large volumes of unstructured text from various sources through the following techniques:

- NLP uses the power of machine learning to remove noise from gathered data – for example, it removes extra whitespaces and other unneeded HTML tags from scraped web data.
- Extract named entities from collected data, including individuals, companies and organizations' names, events, places and date/time. This allows OSINT researchers to identify key actors in large volumes of data quickly.
- NLP helps categorize and classify information according to different criteria based on the research objectives.
- Links between different entities existed in collected data. For example, NLP algorithms can link between the identified entities and events or incidents in the gathered data.

Examples of NLP tools:

- Monkeylearn (<https://monkeylearn.com/sentiment-analysis-online>)
- Lexalytics (<https://www.lexalytics.com>)
- Sentigem (<http://sentigem.com/#!>) (see Figure 7.5).

Figure 7.5: Sentigem (sentiment analysis engine).



Perform a social network analysis

A social network analysis will identify key influencers within a set of data collected from social media websites. This analysis allows for the uncovering

of hidden relationships between different entities. For example, social network analysis helps map out individual relationships associated with a particular group, company, organization or event.

Here are some tools for assisting OSINT researchers in social network analysis:

- NodeXL (<https://nodexl.com>)
- ORA Pro (<https://netanomics.com>).

Image and video analysis

During the analysis phase, OSINT researchers will encounter many cases where they want to analyze specific images or video files – for example, searching for a particular face that appears in a photo or video or analyzing surveillance footage to identify people, locations, or activities.

There are numerous tools for conducting image and video analysis, such as:

- Google Cloud Vision (<https://cloud.google.com/vision?hl=en>) for image and video analysis
- Amazon Rekognition (<https://aws.amazon.com/rekognition>) for image and video analysis
- OpenCV (https://huggingface.co/spaces/AnkitGaur2811/Image_Conversion_app_using_OpenCV)
- Bing Visual Search (<https://www.bing.com/visualsearch>) for executing reverse image search
- Google image search (<https://www.google.com/imghp?hl=en>) for executing reverse image search
- ExifTool (<https://exiftool.org>) for viewing image metadata
- Exif Pilot (<https://www.colorpilot.com/exif.html>) for viewing image metadata.

7.2.4 Phase 4: Reporting

In this phase, OSINT researchers need to convert their finding into a complete report for the requesters. The report will include key findings, recommendations, and any further actions that should be performed.

The report should be written in a way that is easy for non-technical people to digest. For instance, if the OSINT case involves investigating a cybercrime case, the technological terms should be simplified and described clearly to avoid misunderstanding by non-technical people. Here are some tips for creating an efficient OSINT report:

- Categorize your findings into groups and create a key takeaways section at the top of the report to summarize your findings.

- Decide the reporting format – for example, an MS Word file or a presentation using MS PowerPoint, or a combination of both.
- Ensure to cite the sources of your information to ensure credibility.
- If you will share the report with remote stakeholders, encrypt it properly before sending it, especially if the report contains confidential information.

We need to use visual diagrams, charts and other drawing software to showcase some findings. Here are some tools for use in the reporting phase:

- MS Office for report writing – you can use Google docs as an alternative
- Canva (<https://www.canva.com>) – for creating visual reports
- Tableau (<https://www.tableau.com>) – for data visualization
- MS Power BI (<https://www.microsoft.com/en-us/power-platform/products/power-bi>) – for business intelligence and data visualization
- Gephi – for network visualization.

After delivering the final OSINT report, it is critical to stay up to date with the latest trends – especially in the field of advanced internet searching and how to leverage AI tools to aid in your OSINT research. Some OSINT researchers consider the ongoing training as the last phase in the OSINT gathering plan.

7.3 Chapter Summary

The OSINT gathering process or life cycle involves four main phases, starting with preparation and setup, where the search scope is defined and necessary tools are configured, such as VPN and AI-powered search tools in addition to required software to access the darknet such as TOR and I2P. Next is data collection, utilizing techniques like web scraping, social media monitoring, and data mining to harvest data from different online sources. The third phase focuses on data analysis, leveraging natural language processing, social network analysis, and image/video analysis techniques. The fourth and final phase involves reporting and generating comprehensive reports to communicate your findings to decision makers. Finally, continuous improvement is crucial to staying updated with evolving technologies and best practices in the OSINT domain.

Further Reading

1. Maltego, "Everything About Social Media Intelligence (SOCMINT) and Investigations" <https://www.maltego.com/blog/everything-about-social-media-intelligence-socmint-and-investigations> Accessed 2024-05-08

2. Sociallinks, "Top Social Media Intelligence (SOCMINT) Tools in 2023" <https://blog.sociallinks.io/top-social-media-intelligence-socmint-tools-in-2023> Accessed 2024-05-08
3. Researchgate, "Introducing social media intelligence (SOCMINT)" https://www.researchgate.net/publication/262869934_Introducing_social_media_intelligence_SOCMINT Accessed 2024-05-08



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

CHAPTER

8

The Web Layers: Introduction to Surface, Deep and Darknet

How well do you know the internet? Socializing on social media platforms like Facebook and Twitter, watching YouTube videos, posting to the Reddit platform, reading popular blogs and news websites. Conducting searches using popular search engines like Google and Bing will not make you a professional internet user, as everything you see on these sites is simply a part of the surface web which only constitutes 4% of the entire web content!

Cyberspace is huge and it is bigger than what we see when surfing the internet and doing our regular browsing tasks. No one can estimate the exact size of the web, nevertheless the continual adoption of IT in all aspects of life, in addition to the increase in the number of internet users, day after day will certainly increase web content to unpredictable rates.

As we already mentioned, most internet users only access the surface web when doing their regular online tasks, the surface web is the portion of the web that typical search engines can access and index its contents. The second layer of the web is the deep web, this layer is the largest one in size and contains within it another hidden sub-layer which is the dark web – or darknet. Surface and deep web can be accessed using a regular web browser like Firefox and Chrome; however, things are not the same with the darknet, which needs special software to access it.

In this chapter, we discuss different layers that form the web and describe what we expect to see in each layer. However, before we begin talking about the web layers, readers should be able to differentiate between two terms that most internet users use interchangeably which are: The World Wide Web

(WWW) and the internet. For instance, the internet is the network and the IT infrastructure used to access contents on the Web, while the Web is the collection of information – webpages – accessed via the internet.

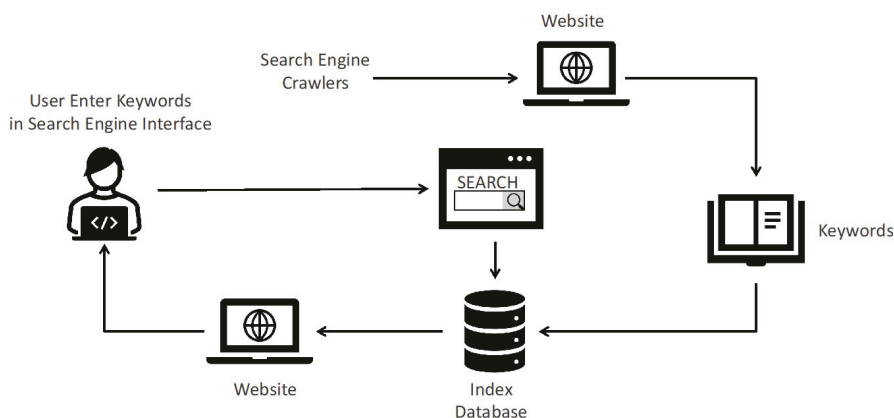
8.1 Surface Web

Also known as the visible or clear web, the surface web is the portion of the web that can be indexed and accessed using standard search engines like Google, Yahoo! and Bing, it constitutes about 4% of web content and can be accessed using standard web browsers (Microsoft Edge, Firefox, Google Chrome) without using any software or special configurations.

Regular search engines index web content by sending robots (also known as crawlers or spiders) to discover new and updated web content. These crawlers travel across the internet and discover new content by following hyperlinks in the visited domain name. For example, when the crawler visits the home page of <https://www.cloudsecasia.com>, it will click – and follow – all hyperlinks on the home page and add the URL of each discovered page to the search engine index database.

When a user wants to use a search engine to look up something online, they need to supply a search query (or keywords); now the search engine looks up the searcher's query in the index database and fetches the results accordingly, starting from the most relevant and ending with the least (see Figure 8.1).

Figure 8.1: How search engines work.



8.2 Deep Web

This layer constitutes the most significant portion of web content (about 96%). It contains all contents that standard search engines cannot index for different reasons, such as content hidden behind login forms that need credentials to access, public databases that require a user to supply a search query to retrieve information from it, fee-based content – like digital libraries, some online magazines, and news channels –that require registration and paying a fee to access in addition to some file types that search engines cannot index.

You can see a list of file formats that the Google Search can crawl, index, and search at: <https://developers.google.com/search/docs/crawling-indexing/indexable-file-types>

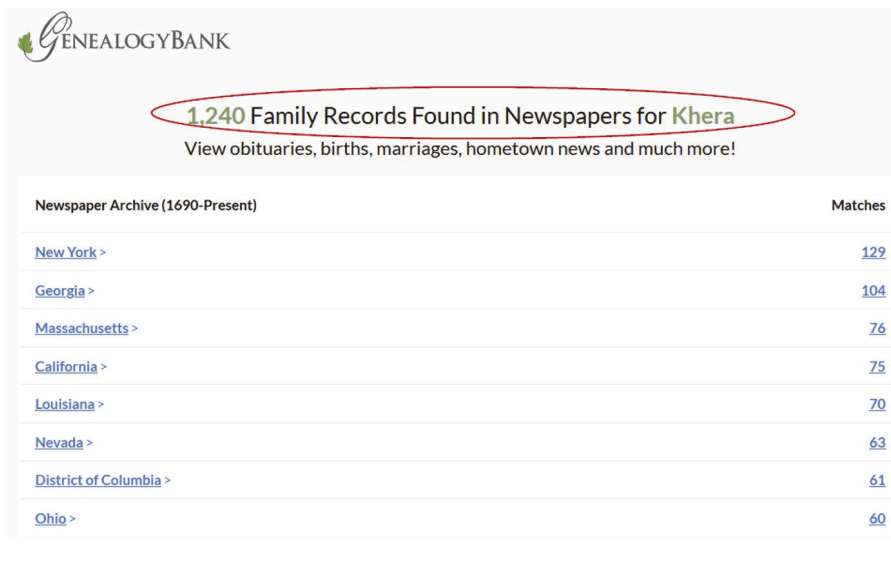
An example of deep web content stored in buried online databases is the *GenealogyBank* (<https://www.genealogybank.com>). To retrieve information from this database, you must supply the family's last name and hit the “Begin Search” button (see Figure 8.2).

Figure 8.2: Searching the GenealogyBank database.

Now, the website's internal search function will search its database and return relevant results (see Figure 8.3). Standard search engines cannot index

online database content because search engine spiders are designed to follow hyperlinks and not to enter search queries and submit search forms to extract results.

Figure 8.3: Sample result page retrieved from a deep web database.



The screenshot shows the GENEALOGYBANK website. At the top, the logo and name 'GENEALOGYBANK' are visible. Below this, a red oval highlights the text '1,240 Family Records Found in Newspapers for Khera'. Underneath this, a subtitle reads 'View obituaries, births, marriages, hometown news and much more!'. The main content is a table with two columns: 'Newspaper Archive (1690-Present)' and 'Matches'. The table lists various US states and the District of Columbia, each with a corresponding number of matches. The states listed are New York, Georgia, Massachusetts, California, Louisiana, Nevada, District of Columbia, and Ohio.

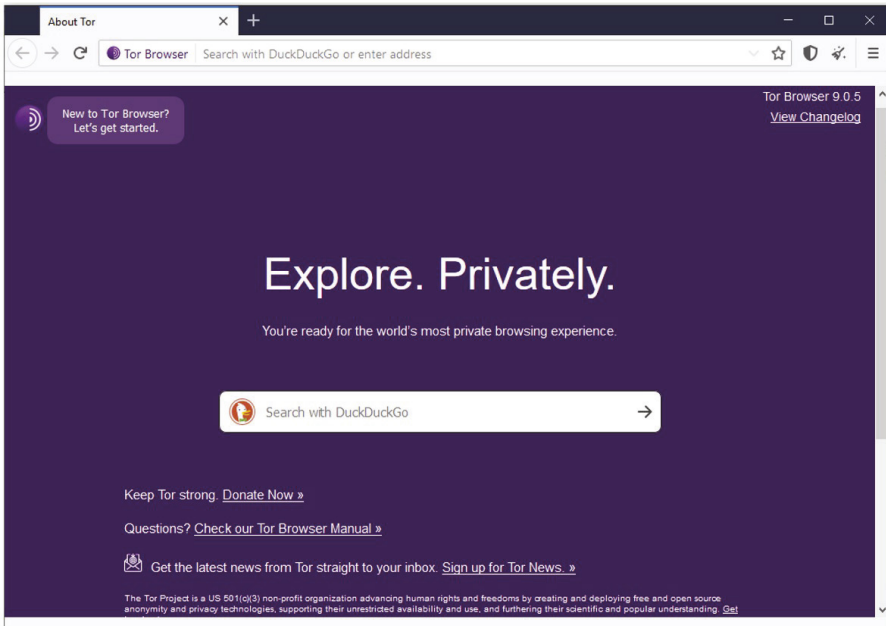
Newspaper Archive (1690-Present)	Matches
New York >	129
Georgia >	104
Massachusetts >	76
California >	75
Louisiana >	70
Nevada >	63
District of Columbia >	61
Ohio >	60

8.3 Darknet

The dark web is a subset of the deep web; it is a collection of private networks (darknet) that constitute what is known as the dark web. We cannot access dark web sites using regular web browsers, as they need special software such as the TOR browser to access (see Figure 8.4). Besides the access, dark web content is encrypted and cannot be indexed using conventional search engines, this makes browsing darknet content relatively difficult compared to surface net. The darknet name is usually associated with illegal and criminal activities; however, a good portion of it is used for noble purposes, e.g., human rights activists and journalists who want to keep their identity and online communications anonymous. The most popular darknet networks are TOR (<https://www.torproject.org/download>), I2P (<https://geti2p.net/en>) and the Freenet (<https://freenet.org>).

No one knows volume of the dark web as there is no way to index its content. However, a study by the *Recorded Future* company⁸ found about 55,828 different onion domains on the TOR darknet (TOR websites use the *.onion* extension as a top-level domain (TLD)).

Figure 8.4: TOR browser used to access the TOR darknet.



8.4 Chapter Summary

Understanding how to mine information from the deep and dark web becomes an essential skill for any cybersecurity professional working to protect IT systems in today's digital age.

Now that we know what the surface, deep, and dark web mean, it is time to dive deeper and see how the darknet can be accessed and searched, and this is what we are going to cover in the next chapter.

⁸Cyberscoop, "How many dark web marketplaces actually exist? About 100." <https://www.cyberscoop.com/dark-web-marketplaces-research-recorded-future/> Accessed 2024-05-01

Further Reading

1. Google Developers, "In-depth guide to how Google Search works" <https://developers.google.com/search/docs/fundamentals/how-search-works> Accessed 2024-05-08
2. St. Louis Community College Libraries, "The Deep Web Library Guide" <https://guides.stlcc.edu/deepweb/welcome> Accessed 2024-05-01

CHAPTER

9

Darkweb and Internet Anonymity: Exploring the Hidden Internet

As mentioned in the previous chapter, the dark web is a tiny part of the deep web that we cannot access using regular search engines. The dark web is not one single network; it is composed of many private darknets that collectively form what is known as the dark web. A darknet network is a decentralized peer-to-peer network. Some of the darknet networks are relatively large and powered by official organizations or public initiatives, while others are small and run by individuals. The most famous dark networks are TOR (<https://www.torproject.org>), I2P (<https://geti2p.net>), Freenet (<https://freenetproject.org/index.html>), and Riffle (<https://github.com/kwonAlbert/riffle>).

To access darknet websites, special software is needed, as darknet networks are encrypted and cannot be accessed directly via regular web browsers. For example, you should use the TOR browser to access the TOR network.

The dark web is famous for hosting websites that promote illegal products and services (see Figure 9.1), such as the drugs and arms trade, stolen financial and private data, false government documents and more. Commercial transactions in this hidden layer occur via cryptocurrencies (mainly Bitcoin). The vast spread of unlawful activities should not prevent us from accessing the legal side of this layer. For instance, darknets contain many legal websites commonly run by human rights activists and journalists that exploit the anonymity of this network to promote their activities while remaining anonymous online.

In this chapter, we discuss the dark web, how to access it, and where to start searching it. TOR is the most popular anonymity network in the dark web, so it will be the focus of this article.

Figure 9.1: Illegal TOR service that promotes selling false government documents.



9.1 TOR Network

The TOR anonymity network is free, open source software originally developed by the US Navy Research Laboratory (NRL) in the mid-1990s to protect intelligence online communications. Later, in 2004, the NRL released the TOR source code under a free license. TOR is now managed by *The Tor Project, Inc*, a non-profit organization that maintains TOR development.

TOR network is composed of two components:

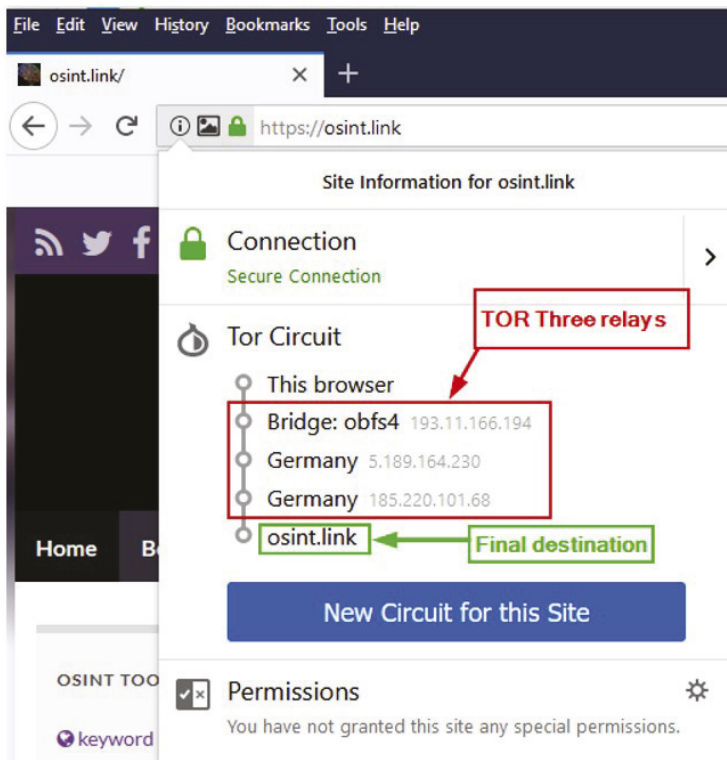
1. The TOR software used to access the TOR network.
2. TOR infrastructure is simply a set of volunteer computers spread worldwide that route users' traffic across the TOR network.

TOR network can be used to anonymize an internet user's online identity when surfing the surface web, in addition to enabling access to TOR anonymous websites (also known as TOR hidden services). To use TOR for browsing the surface internet anonymously, all you need to do is to download the TOR browser

(<https://www.torproject.org/download>) and use it as you would use any your regular web browser to surf online.

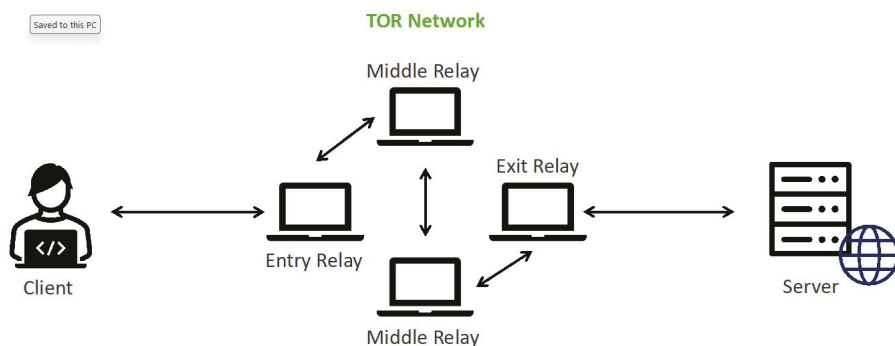
TOR anonymizes internet traffic by bouncing your connection over at least three relays, also known as nodes or routers, before reaching the destination; depicted in Figure 9.2. The first relay is the entry relay, also known as the guard relay; it routes your data from the surface web and moves it to the middle relay. TOR uses at least one middle relay for each connection. However, it may use more; the middle relay then shifts your connection to the final one, the Exit relay; see Figure 9.3. When surfing surface websites using the TOR browser, your actual IP address will be concealed, and your connection will seem to originate from the TOR Exit relay IP address instead. Using TOR will efficiently hide your identity online and prevent websites from tracking your browsing history. TOR is also considered a valuable tool for circumventing censorships implemented in many countries.

Figure 9.2: Example of TOR relays to anonymize the user IP address.



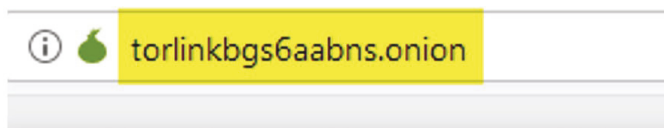
Data going through the TOR network is encrypted; however, when the data leaves the TOR network at the Exit relay, it needs to be decrypted again to continue to the final destination on the surface web. A malicious actor can misconfigure the Exit relay to intercept the traffic, so it is advisable to encrypt your sensitive data before sending it via the TOR network.

Figure 9.3: How TOR anonymizes connections into the surface web.



The address of the TOR hidden services, also known as onion services, is composed of a series of random string characters (numbers and letters only) and ends with the `.onion` extension (pronounced “dot-onion”); see Figure 9.4. A TOR hidden service is fully anonymous and allows its owner to use it to promote various activities on the TOR network, like anonymous instant messaging and web publishing in addition to sharing files anonymously.

Figure 9.4: Sample TOR hidden service URL.



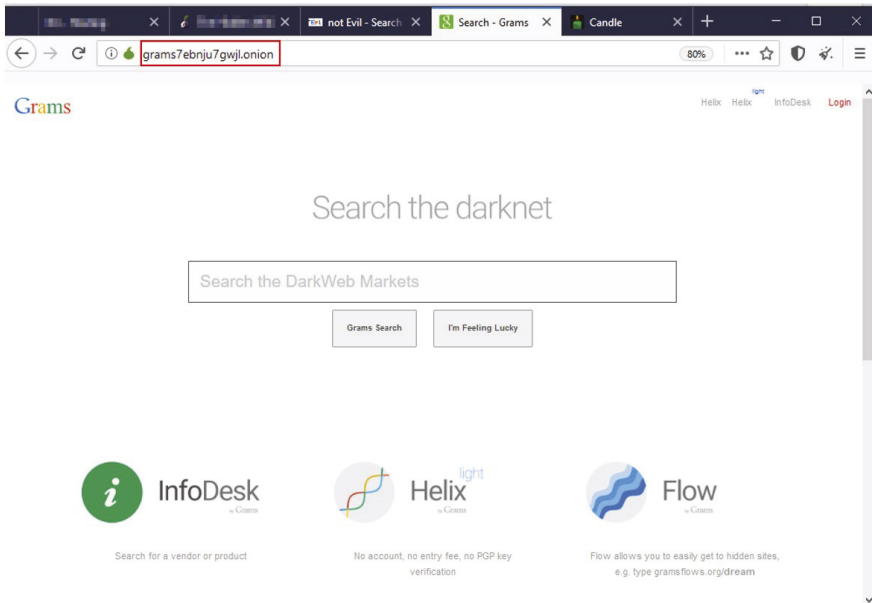
It is a good practice to conceal your entrance to the TOR network, although using TOR is considered legal in most countries, using it may be considered somehow suspicious and raise a red flag, especially when TOR traffic passes through the government’s firewall. To conceal your TOR usage, use a VPN before connecting to the TOR network. TOR also provides a mechanism to hide its usage, known as “pluggable transport”. You can find a full guide on how to use TOR pluggable transport at <https://2019.www.torproject.org/docs/pluggable-transport.html.en>.

9.2 Searching the TOR Network

Surfing the TOR network is not easy as surfing the surface websites. The ephemeral nature of TOR hidden services, in addition to the absence of powerful search engines (like Google and Bing) to index the TOR darknet, makes finding onion sites a problematic task. Many dedicated TOR search engines and TOR portals try to simplify finding TOR hidden services by providing a directory of TOR sites. The following are the most popular TOR search engines/directories:

1. The Hidden Wiki (<http://zqktlwuavvvqqt4ybvvgvi7tyo4hjl5xgfuvpdf6otjiycgwqbybm2qad.onion>)
2. Ahmia (<http://msydgstlz2kzerdg.onion>)
3. Torch (<http://xmh57jrznw6insl.onion>)
4. TorLinks (<http://torlinksg6enmcyuyxjpjkoouw4oorgdgeo7ftnq3zodj7g2zxi3kyd.onion>)
5. NotEvil (<http://hss3uro2hsxfogfq.onion>)
6. Grams (<http://grams7ebnju7gwjl.onion>) (see Figure 9.5)
7. Candle (<http://gjobqjj7wyczbqie.onion>)
8. Tor Onionland (<http://3bbaaacczczbdddz.onion>)

Figure 9.5: Grams TOR search engine



9.3 Chapter Summary

The dark web is a term used to name decentralized peer-to-peer anonymous networks. There are many anonymity networks, and the most popular one is the TOR network.

TOR is famous among the general public because it allows its users to surf the surface web anonymously at an acceptable speed, in addition to the regular function of darknets which is hosting anonymous websites.

Dark web networks cannot be searched or indexed using regular search engines like Google and Yahoo; however, for each darknet, you can find many specialized search engines and directory sites that list popular sites hosted in that darknet.

In the next chapter, we will talk about another important cybersecurity field related to OSINT, i.e., digital forensics. Digital forensics is a field used to investigate digital crimes. It is a relatively emerging field in computer security, and its operations intersect many disciplines, such as online investigations, eDiscovery, and intrusion investigation.

Further Reading

1. Socradar, "Top 5 Dark Web Search Engines" <https://socradar.io/top-5-dark-web-search-engines> Accessed 2024-05-02
2. Avast, "The Dark Web Browser: What Is Tor, Is It Safe, and How to Use It" <https://www.avast.com/c-tor-dark-web-browser> Accessed 2024-05-08

CHAPTER

10

Introduction to Digital Forensics

As the world continues to digitize, the dependence on technology to do business will increase in both the public and private sectors. In today's information age, organizations use technology to increase productivity, save internal and external operational costs, enhance data security, and expand business capabilities. The main key to achieving these benefits is to implement digital transformation of all work aspects, especially by storing data digitally to replace paper files. Individuals also become very dependent on technology in their daily lives. Almost everything people do involves technology in one way or another.

A growing increase in cybercrimes accompanied the rapid shift to the digital age. According to *Cybersecurity Ventures*⁹, cybercrime damages will reach US\$10.5 trillion annually by 2025. The same study predicts that there will be 7.5 billion internet users by 2030. Consequently, the proliferation of digital devices will generate a massive amount of digital data every second.

Digital forensics, also widely known as computer forensics, is the process of investigating crimes committed using any type of computing device, such as computers, servers, laptops, cell phones, tablets, digital camera, networking devices, Internet of Things (IoT) devices or any type of digital data storage device. Digital forensics also covers examining attacks originating from cyberspace like ransomware, phishing, SQL injection attacks, distributed denial-of-service (DDoS) attacks, data breach, and any cyberattacks that cause

⁹Cybersecurity Ventures, "Cybersecurity Ventures Official Annual Cybercrime Report." <https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016> Accessed 2024-05-01

financial or reputation losses. The ultimate goal of a digital forensics investigation is to preserve, identify, acquire and document digital evidence to be used in a court of law.

Under this definition, digital forensics is used to investigate any crime involving electronic devices, whether these devices were used to commit or as a target of a crime. Having a digital forensics capability becomes very important for modern organizations in investigating internal policy violations and external attacks against their computerized systems. Large corporations already have capabilities that exceed the capabilities of many government police departments.

10.1 Digital Evidence

As already mentioned, the main task of any digital forensics investigation is to acquire, preserve, examine and present digital evidence to be used in a court of law, so what is meant by the term “digital evidence”?

Digital evidence, also known as electronic evidence, is any information stored or transmitted in digital format. This includes data found on computers, laptops, cell phones, tablets, PDA hard drives, and all data stored using various storage device media such as USB thumb drives, SD cards, external hard drives, and CD/DVD. Data transmitted via computer networks is also considered a part of digital evidence in addition to operating systems and database logs.

Digital evidence should be acquired in a forensically sound manner. "Forensically sound" is a term used by digital forensics examiners to describe the process of acquiring digital evidence while preserving its integrity to be admissible in a court of law.

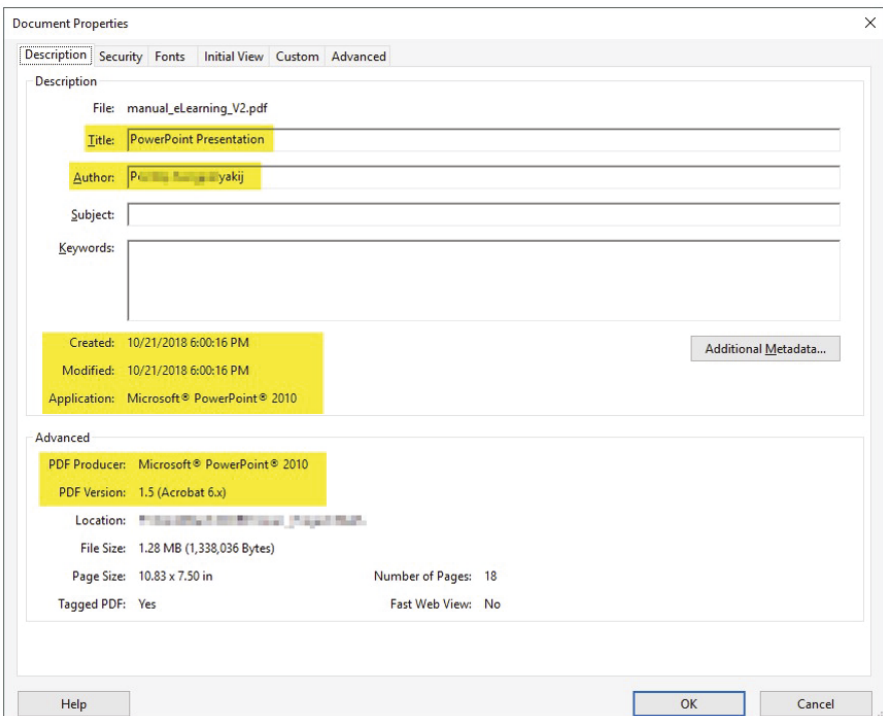
Digital evidence includes the following artifacts – and more:

1. Email messages and attachments
2. User account info (username, password, personal picture, etc.) for both online accounts (cloud storage, social media) and local computer accounts
3. Digital photo, audio, and video
4. IM conversation history
5. Web browsing history
6. Files generated by accounting programs
7. All types of electronic files (MS Office files, databases, spreadsheets, bookmarks, etc.)
8. Data in volatile memory (RAM)
9. Registry info (in Windows-based systems)

10. Computer backups
11. Networking device logs (router, switch, proxy server, firewall)
12. Printer spooler files
13. ATM transaction logs
14. Fax and copier machine logs
15. Electronic door locks logs
16. GPS track logs
17. Digital data extracted from home appliances (smart TV, smart refrigerator)
18. Surveillance video recordings
19. Encrypted and hidden files
20. And any data stored in digital format and can be used in the court of law as evidence.

We should take note that most digital file types have associated metadata, as discussed in Chapter 5 and shown in Figure 10.1. For instance, metadata is data about data, some are automatically generated by the application that created

Figure 10.1: Sample metadata found in a PDF file.



the file, e.g., file creation and alteration date/time, application version info, and the users themselves can set other metadata, e.g. the file's author name, comments, and email address.

10.2 Digital Forensics Process

There are many methodologies or suggested processes for conducting digital forensics investigations; however, they all share the following four key main phases, see Figure 10.2.

Figure 10.2: Common phases of digital forensics.



10.2.1 Seizure

In this phase, the suspect digital device is seized, packaged properly, and taken to the digital forensics lab. An investigator should have an official search warrant from a court, and they must have the proper permission to confiscate the suspect's device.

The digital device can be any type of computing device such as a desktop computer, server, laptop, tablet, smartphone, external hard drive, USB stick or backup media, or Internet of Things (IoT) device.

Upon arriving at the crime scene, if the suspect device is still running, volatile memory should be acquired first (if possible) before powering off the device. Volatile memory can contain important information for the current investigation, such as passwords, IM chat log, internet browsing history, running programs, and clipboard content.

In some cases, jurisdictional challenges may arise and prevent forensic investigators from seizing suspect digital devices. For example, if the crime was committed through the internet and the suspected machine or server was located in another country, how can evidence be acquired in such a case? Having

an international search warrant is difficult and time-consuming and may not be applicable in all cases.

10.2.2 Evidence acquisition

In this phase, a professional computer forensics technician will duplicate the suspect device hard drive – and RAM if applicable – to acquire a complete image of it (also known as bit-to-bit image). Having more than one forensic image is preferable as the examination will be performed on these copies in the lab.

10.2.3 Evidence analysis

In this phase, the acquired forensic image is analyzed using different tools and techniques to acquire useful leads from it, such as recovering deleted files and emails, discovering hidden data, retrieving IM chats and web browser history. The forensic tool/s used to analyze the forensic image should be accepted by a court of law. Some popular court-accepted tools include the following: EnCase (<https://www.guidancesoftware.com/encase-forensic>), Sleuth Kit (<https://www.sleuthkit.org/autopsy>), Volatility (<https://www.volatilityfoundation.org>) for capturing RAM memory, and AccessData (<https://accessdata.com>).

10.2.4 Evidence presentation

In the final phase, the forensic investigator will produce a comprehensive report that details his/her findings. The language used to write the report should be well understood by non-technical people such as attorneys, judges, and juries.

10.3 Digital Investigation Types

Digital forensic investigations can be categorized into two categories according to who initiated the investigation:

1. *Public investigations* include all criminal cases managed by government law enforcement agencies and are conducted according to country law. This type of investigation follows three main stages: complaint, investigation, and prosecution.

2. *Private investigations*: This type of investigation is conducted by corporations to investigate different cases related to computer crime targeting their IT systems, such as policy violations, examining wrongful termination, or leaking of enterprise secrets. No formal rules govern such cases, as each organization has its own rules. Nevertheless, private investigations should be conducted following the same strict procedures as public investigations, as private investigations can later move to the court and become official criminal cases.

10.4 Digital Forensics Readiness

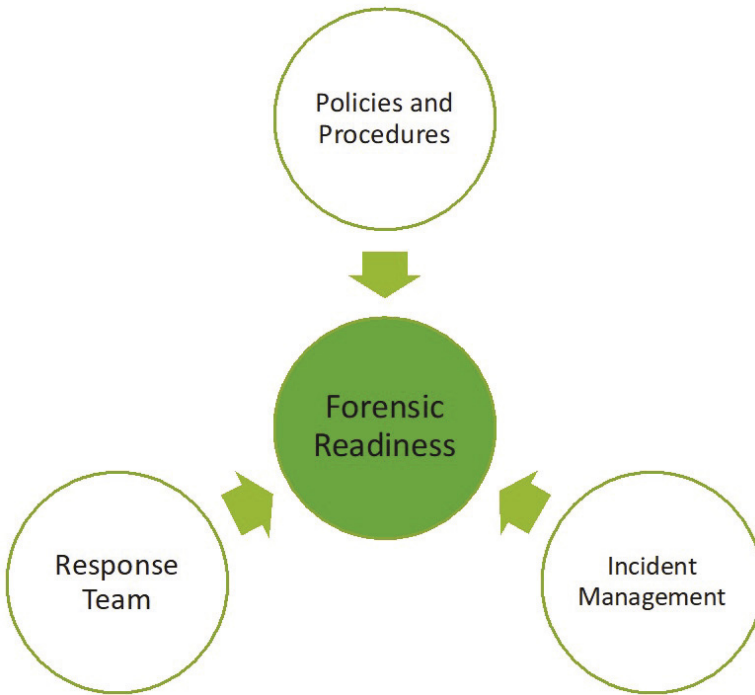
As more organizations become – almost – fully dependent on technology to run their business, having a sudden standstill caused by an unwanted incident can have catastrophic consequences on their operations. For this reason, strategies like disaster recovery, incident response, and digital forensic investigation should be fully incorporated into the organization's operational structure.

An organization's forensics readiness, see Figure 10.3, is defined as its ability to collect, preserve, protect, and analyze digital evidence in a forensically sound manner whenever an incident occurs. This will help it reduce downtime, adequately investigate the criminal case, and shift it to court if necessary. Having a forensic readiness plan brings many advantages for organizations, such as:

1. Force employees to avoid violating company policy as they will have the sense that they will get caught easily if they carry out any illegal activity.
2. Forensic readiness will increase the organization's ability to discover cyberattacks against its IT infrastructure before they escalate and become more harmful.
3. Reduce costs associated with digital forensics investigations as the organization will already have the plan, procedures, and tools for acquiring and analyzing digital evidence. This will lead to a fast resolution of any criminal case.
4. Compliance with government regulations: Having a forensic readiness plan becomes mandatory in many countries to ensure an organization's ability to acquire digital evidence in a forensically sound manner when required by an internal investigation and before moving the case to official courts.

Maintaining a forensic readiness plan has become necessary for any organization or corporation that wants to survive in today's information age.

Figure 10.3: Digital forensic readiness.



10.5 Chapter Summary

In this chapter, we introduced the term “digital forensics” and covered the concept of digital evidence, its types, and where we can find it in electronic devices. There is no standard process for conducting digital forensics investigations; however, we introduced the general phases of any digital investigation process and what tasks are required as a part of each phase.

Digital forensics investigators are needed in all business sectors. Most organizations now maintain digital forensics readiness policy to meet government regulatory requirements and to better respond to incidents threatening their business continuity.

In the next chapter, we will cover how to utilize open source intelligence (OSINT) techniques to support digital forensics investigations and locate information about individuals and companies online.

Further Reading

1. Bluevoyant, "What Is Digital Forensics?" <https://www.bluevoyant.com/knowledge-center/understanding-digital-forensics-process-techniques-and-tools> Accessed 2024-05-08
2. Interpol, "GUIDELINES FOR DIGITAL FORENSICS FIRST RESPONDERS" https://www.interpol.int/content/download/16243/file/Guidelines_to_Digital_Forensics_First_Responders_V7.pdf Accessed 2024-05-03

CHAPTER

11

OSINT for Digital Forensics Investigations

The proliferation of the internet has intensified the number of people using social media sites to interact with others and share personal information. This has created a wealth of personal information about almost anyone on Earth that is easily and publicly accessible.

The internet is designed as a public network; this makes anything published online readily available to anyone who has a computer with an internet connection. Although social networking sites allow users to post information privately, the majority of users still share their social activities publicly, making such information freely, legally accessible, and considered a part of OSINT sources.

Traditional digital forensics methodologies are concerned with acquiring digital evidence from physical computing devices; however, with the increased number of social networking users, information available on social platforms and other public online databases should be exploited in any digital forensic investigation that requires using online public sources to obtain information.

In this chapter, we will examine using OSINT tools and techniques as a digital forensic investigative tool, focusing on harvesting data about individuals.

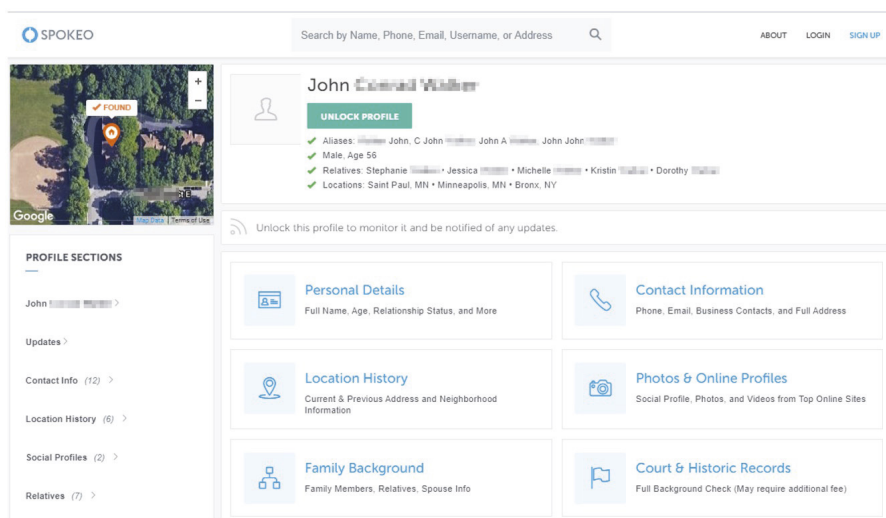
11.1 OSINT to Collect Individual Intelligence

We can search for people using their full name, email address, phone number or mailing address. Many specialized people's data collection websites offer such a service. The following are some famous people search engines:

1. Spokeo (<https://www.spokeo.com>)
2. Thatsthem (<https://thatsthem.com>)
3. Lullar (<https://www.lullar.com>)
4. Zabasearch (<https://www.zabasearch.com>)
5. Truepeoplesearch (<https://www.truepeoplesearch.com>)
6. Truthfinder (<https://www.truthfinder.com>)
7. 411 (<https://www.411.com>).

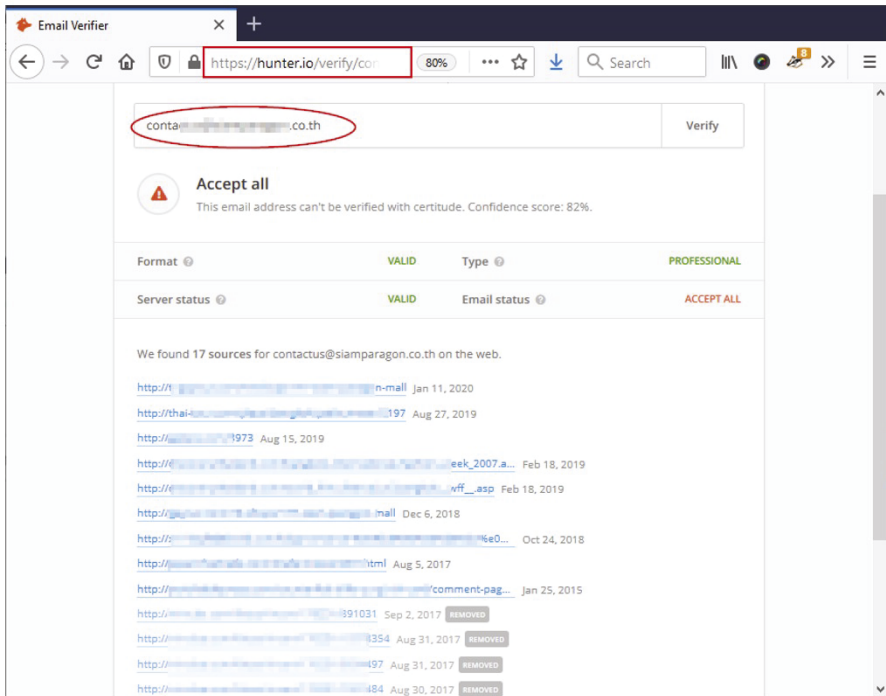
We will experiment with searching for someone using their full name with Spokeo.com (see Figure 11.1) and then using their email address.

Figure 11.1: Using Spokeo to look up people's information.



To look for someone using their email address, we can use hunter.io (see Figure 11.2); however, before starting to search by email, we should check whether the target email address is valid or not. I will use a free email verification service, *Email Dossier* (<https://centralops.net/co/EmailDossier.aspx>) (see Figure 11.3).

Figure 11.2: We can search for someone's email address and see where it appeared online.



To look up the target on LinkedIn using their email address (you should sign in to your LinkedIn account first to view the results), replace TargetEmail@email.com with your target email address in the following URL <https://www.linkedin.com/sales/gmail/profile/viewByEmail/TargetEmail@email.com>

If the target was using that email address for his LinkedIn account and the privacy settings of the target profile are set to allow public search by email, you will see the target profile URL (see Figure 11.4).

If you have the full name of someone and want to see what information is available about them on LinkedIn (or any other site), use the Google Dork in Figure 11.5.

Now that you know the target LinkedIn, you can get their photo LinkedIn profile photo (if the target included a profile photo) and conduct a reverse image search. Reverse image search engines search the web using a picture

Figure 11.3: Using Email Dossier to verify the subject email address.

Email Dossier Investigate email addresses

email address

user: anonymous [IP: ...7.186]
balance: 49 units
[log in](#) | [account info](#) *CentralOps.net*

Validating ...@gmail.com...

Validation results

confidence rating: **3 - SMTP**
The email address passed this level of validation without an error. However, it is not guaranteed to be a good address. [more info](#)

canonical address: ...@gmail.com>

MX records

preference	exchange	IP address (if included)
5	gmail-smtp-in.l.google.com	...194.78....
10	gmail-smtp-...google.com	...253.111...
20	gmail-smtp-...google.com	...194.77....
30	gmail-smtp-...google.com	...33.177....
40	gmail-smtp-...google.com	...194.201...

SMTP session

```
[Contacting gmail-smtp-...oogle.com [1... 26]...]
[Connected]
220 mx.google.com ESMTP q7si...33otk.77 - gmsmtp
EHLO mx1.validemail.com
250-mx.google.com at your service, [IP: ... 247]
250-SIZE 1...6400
250-8BITMIME
250-STARTTLS
```

rather than keywords. The following are some popular reverse image search engines:

1. Yandex (<https://yandex.com/images>)
2. Google reverse search (<https://www.google.com/imghp>)

Figure 11.4: We can find someone's LinkedIn profile using many methods, including using the LinkedIn internal search function.

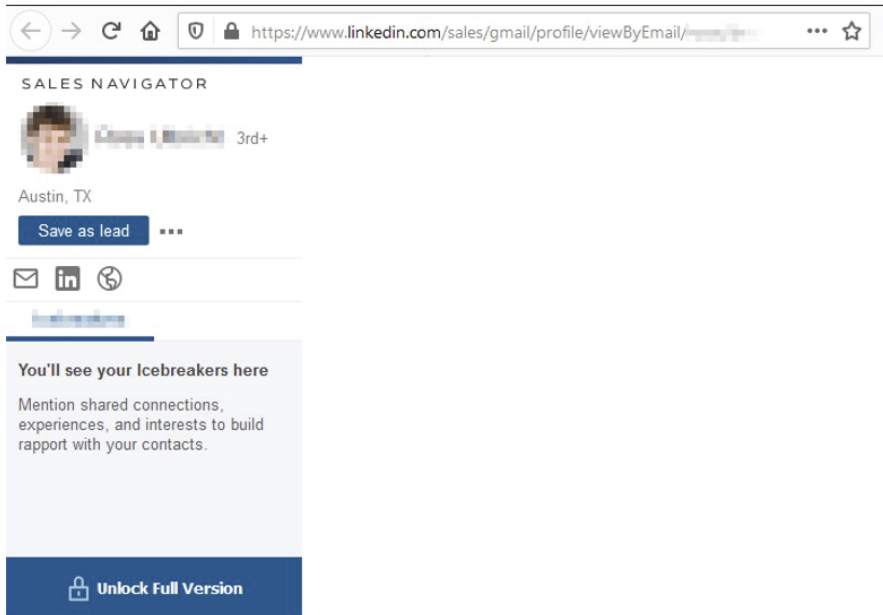
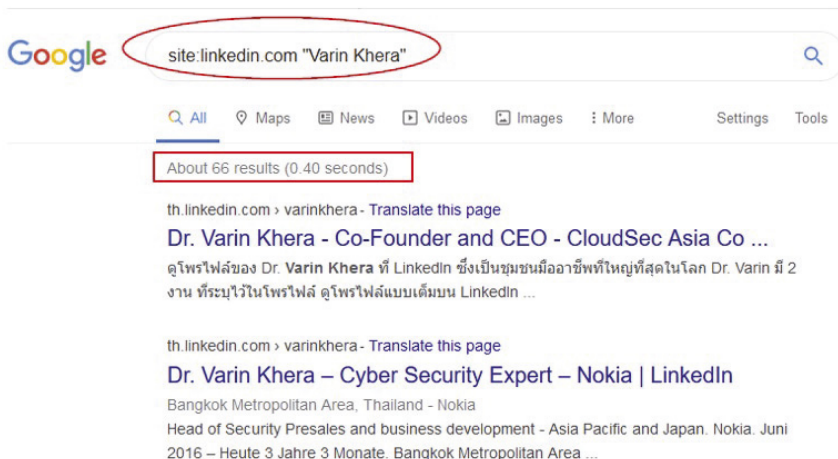


Figure 11.5: Search for a target name using Google – in this example, I restrict the search results to the LinkedIn.com website.

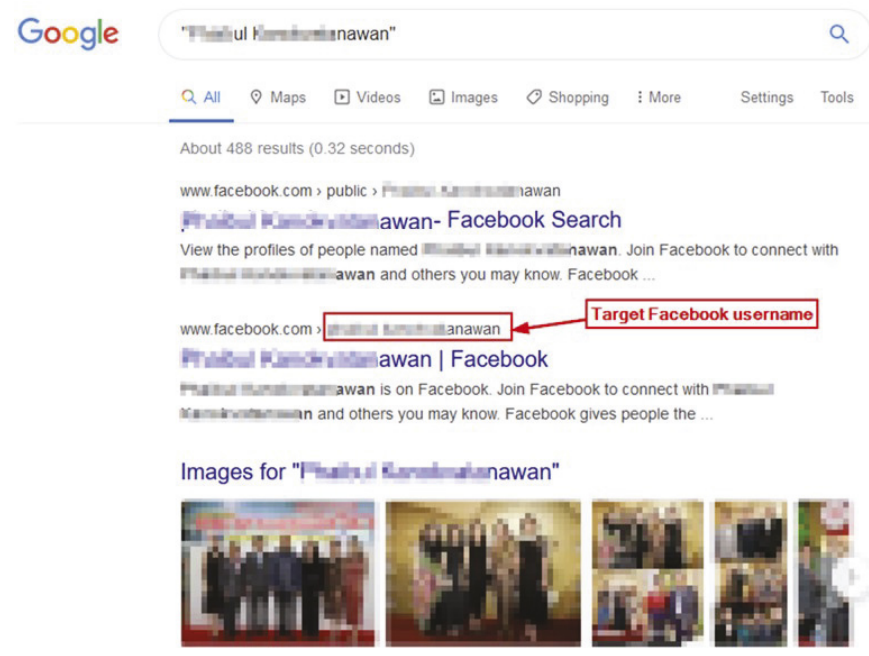


3. TinEye (<https://tinEye.com>)
4. Bing Visual Search (<https://www.bing.com/visualsearch>).

11.2 Investigating Social Networking Sites

Finding the social networking profiles of the target is a great start. By examining the individual's social media activities, we can obtain a lot of helpful information about the target's work, friends, online habits, places previously visited, political and religious opinions and much more. All social media platforms, such as Facebook and Twitter, have a search function where you can search for a target name and examine their public posts for interesting facts. For this experiment, I will use Google to locate one of my target social media profile usernames (e.g., Facebook, LinkedIn, Instagram, Twitter) and then try to see if the target has used the same username in other places (see Figure 11.6).

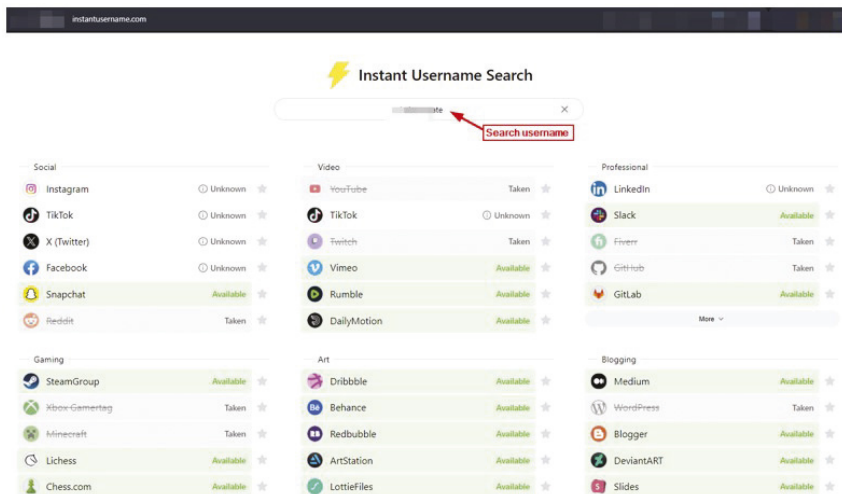
Figure 11.6: Using Google search to locate one social media profile of the target – in my case, I was able to find the target Facebook username.



Now that we have the target Facebook username, we can use an online service to see where else this username is used, as most people prefer to use the same username on multiple social media services. The following websites, search for similar usernames across scores of social media platforms:

1. Namechk (<https://namechk.com>)
2. Instantusername (<https://instantusername.com>) (see Figure 11.7)
3. Usersearch (<https://www.usersearch.org>)

Figure 11.7: Finding all similar social media usernames of the target.



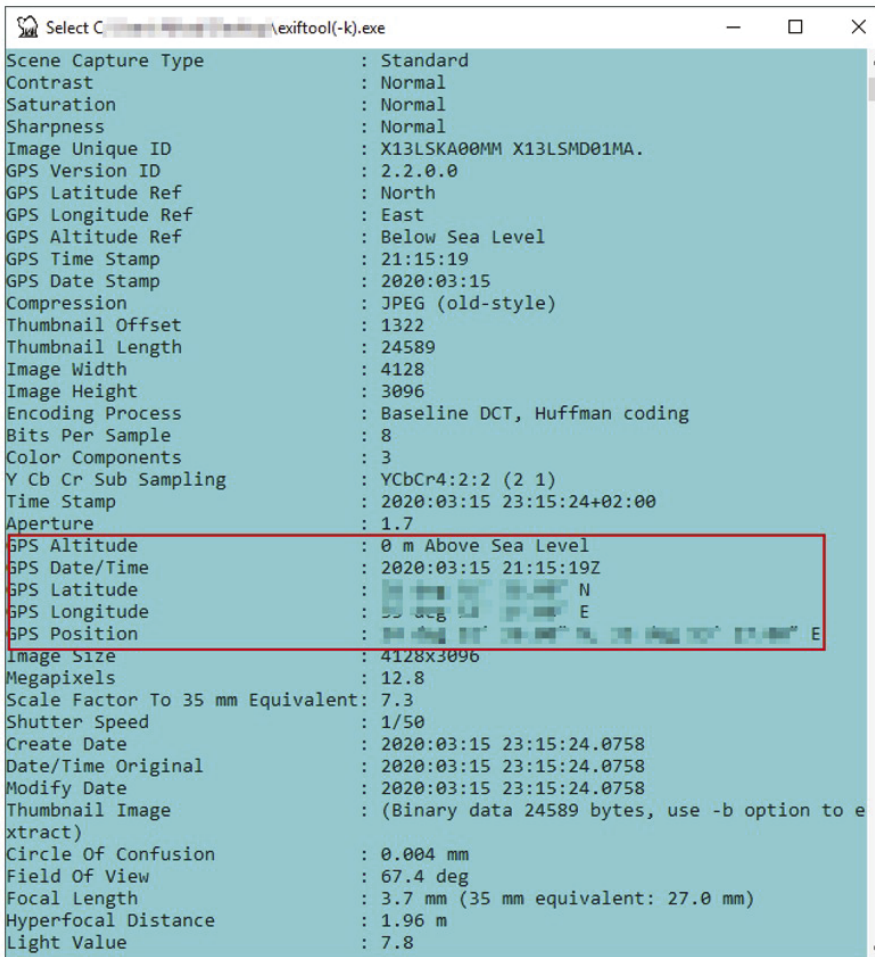
11.3 Investigating a Digital File's Metadata

While searching for someone, you may find media files (images, audio or video) related to the target scattered on different sites. Suppose the target has some academic background or works in a company where they are responsible for posting some types of files online (e.g. Company promotional brochures or job vacancies requirement files). In that case, you can also examine the metadata of these electronic files to obtain helpful information.

Let us experiment with examining a photo metadata for information. We will use the *ExifTool* from *Phil Harvey* (<https://exiftool.org>). To use this tool, just drag the picture from which you want to extract metadata above the program icon, and you will be ready. Figure 11.8 shows the amount of metadata information that can be extracted from a JPG file.

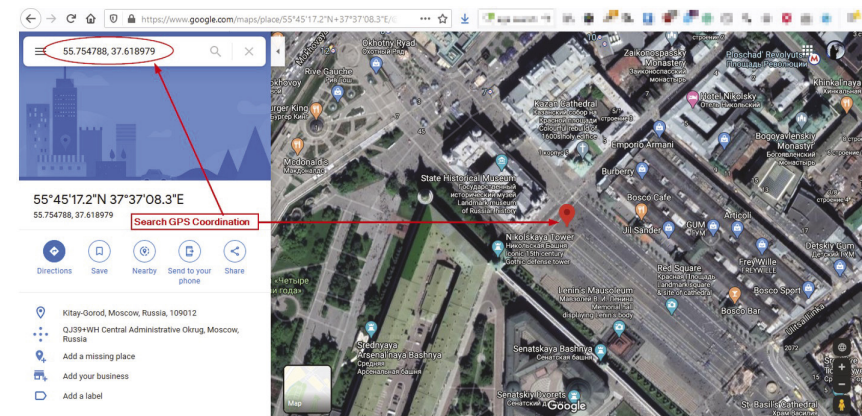
```
C:\Users\... \exiftool(-k).exe
ExifTool Version Number      : 11.10
File Name                    : 20200315_231524.jpg
Directory                    : 
File Size                    : 4.9 MB
File Modification Date/Time   : 2020:03:15 23:43:10+02:00
File Access Date/Time        : 2020:03:15 23:43:10+02:00
File Creation Date/Time      : 2020:03:15 23:43:09+02:00
File Permissions              : rw-rw-rw-
File Type                    : JPEG
File Type Extension          : jpg
MIME Type                    : image/jpeg
Exif Byte Order               : Little-endian (Intel, II)
Make                         : samsung
Camera Model Name             : SM-J730F
Orientation                   : Horizontal (normal)
X Resolution                  : 72
Y Resolution                  : 72
Resolution Unit               : inches
Software                      : J730FXWU4CSF5
Modify Date                   : 2020:03:15 23:15:24
Y Cb Cr Positioning           : Centered
Exposure Time                 : 1/50
F Number                     : 1.7
Exposure Program              : Program AE
ISO                           : 64
Exif Version                  : 0220
Date/Time Original            : 2020:03:15 23:15:24
Create Date                   : 2020:03:15 23:15:24
Components Configuration     : Y, Cb, Cr, -
Shutter Speed Value           : 1/50
Aperture Value                : 1.7
Brightness Value              : 2.47
Exposure Compensation         : 0
Max Aperture Value            : 1.7
Metering Mode                 : Center-weighted average
Flash                         : No Flash
```

Figure 11.8: Files metadata can reveal a great number of details.



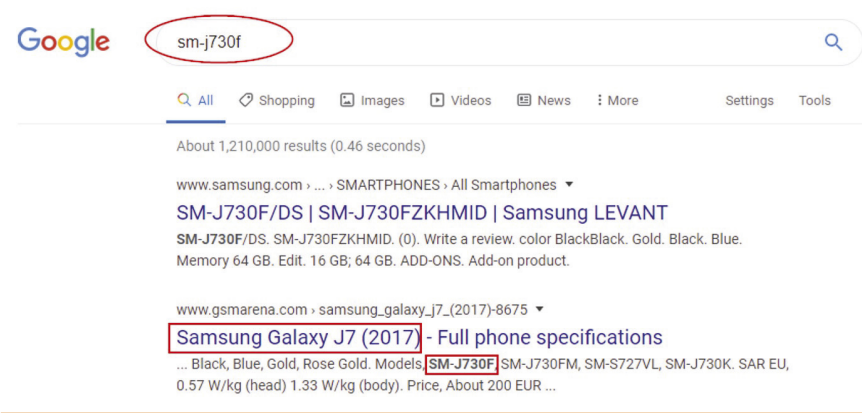
Now that we have the GPS coordination extracted from the image above, we can use Google Maps (<https://www.google.com/maps>) to find the exact geographical location of this photo (see Figure 11.9).

Figure 11.9: Find the geographical location of any place on Earth.



We can also search for the camera model of the capturing device using Google (see Figure 11.10).

Figure 11.10: Searching for the camera model.

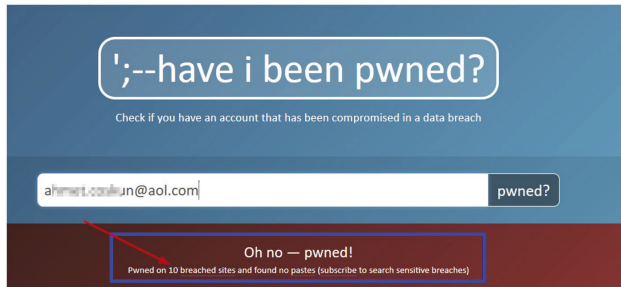


11.4 Searching for Leaked Credentials on the Darknet

Now that we have the target email address and are sure it is valid and working, we can search for breached accounts with the associated target email. For instance, many online repositories list leaked credentials from

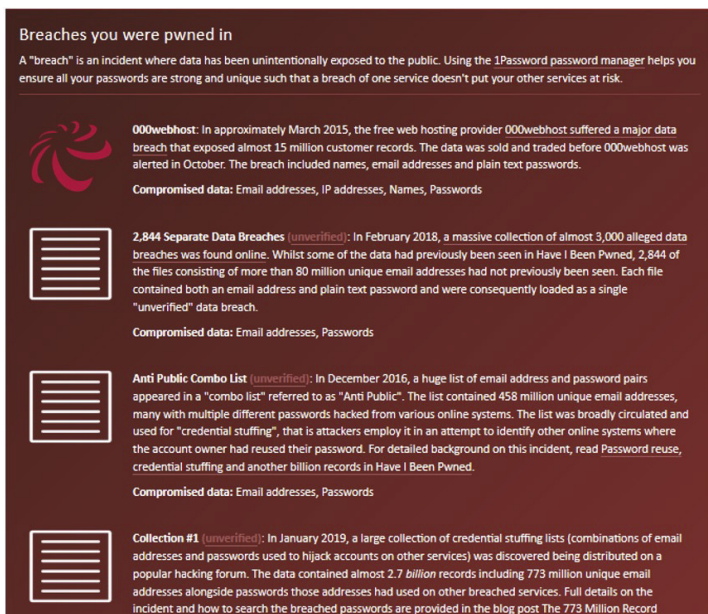
breached websites. For this experiment, we will use *have i been pwned?* (<https://haveibeenpwned.com>) (see Figure 11.11).

Figure 11.11: Searching for subject email address reveal related accounts breached in 10 websites.



For each breached website found, *have i been pwned?* will list detailed information about each breach (see Figure 11.12).

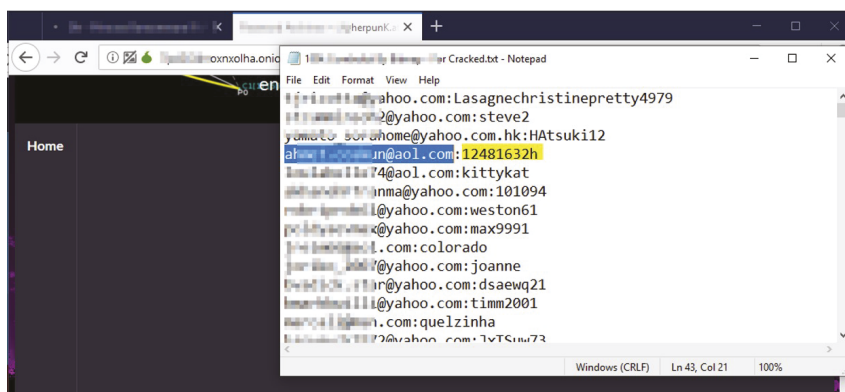
Figure 11.12: Have i been pwned? lists detailed information about each breach.



Now, we have many options to find the credentials associated with each breached account. For instance, we will search the darknet to see if we can find something about the target email address.

Launch TOR browser and begin your search. We were lucky to find this file that lists the associated password of one of the target email breached accounts (see Figure 11.13).

Figure 11.13: Find leaked credentials on the darknet.



11.5 Chapter Summary

OSINT techniques can be used to harvest information about any entity (whether an enterprise or individual) online to support our digital forensics investigation. This chapter gives practical scenarios for exploiting public data sources by using different OSINT techniques to obtain valuable intelligence about our targets. In the next chapter, we will conclude this book by discussing relevant cyber security and data protection laws in the Asia Pacific region.

Further Reading

1. Cellebrite, "Unlocking the Power of Open Source Intelligence (OSINT) in Digital Forensics" <https://cellebrite.com/en/unlocking-the-power-of-open-source-intelligence-osint-in-digital-forensics> Accessed 2024-05-02
2. Cobwebs, "Deepening the Bond Between OSINT and Digital Forensics" <https://cobwebs.com/en/blog/osint-digital-forensics-partnership> Accessed 2024-05-03

CHAPTER

12

Data Protection and Cybersecurity Laws for the Asia Pacific Region

The number of people using the internet is increasingly growing, with more than one million users accessing the internet for the first time each day¹⁰. *Cybersecurity Ventures* predicts there will be 7.5 billion internet users by 2030 (90% of the projected world population of 8.5 billion, six years of age and older)¹¹.

Aside from commerce sales, most internet users access it to socialize and interact with their peers online. For instance, there were 3.80 billion social network users in January 2020, which had increased by about 9% since the previous year¹².

The huge transformation from the real world to the digital universe will make most people, enterprises, and government interactions happen in cyberspace. The advance of the internet and related communications technology allows easy access to information from anywhere on earth. For example, an online merchant operating in Thailand can offer services to customers in the EU and the USA. To handle the spread of personal, financial, medical

¹⁰Clickz, “Internet growth + usage stats 2019: Time online, devices, users” <https://www.clickz.com/internet-growth-usage-stats-2019-time-online-devices-users/235102> Accessed 2024-05-01

¹¹Cybersecurity Ventures, “Humans On The Internet Will Triple From 2015 To 2022 And Hit 6 Billion” <https://cybersecurityventures.com/how-many-internet-users-will-the-world-have-in-2022-and-in-2030> Accessed 2024-05-01

¹²edubirdie, “Digital in 2020: Transformative Impact and Key Highlights” <https://edubirdie.com/blog/3-8-billion-people-use-social-media> Accessed 2024-05-01

and other types of personal information across the globe via the internet, the appropriate legal regulations should be set to protect citizens' private data and organizations' digital assets when working online.

Following the implementation of the General Data Protection Regulation (GDPR) in the European Union (which came into force on 25 May 2018), which regulates data protection and privacy in EU countries in addition to the transfer of personal data outside the EU and EEA areas, more countries in the world began to review and strengthen their data protection and cybersecurity laws to cope with the new regulation. While the GDPR is an EU regulation, enterprises operating outside EU countries should be aware of its implications to avoid violating any of its terms when dealing with or processing the personal data of EU citizens.

In the final chapter of this book, we will briefly review the cybersecurity and personal data protection acts implemented in major countries in the Asia Pacific region. Keep in mind that cybersecurity and internet privacy laws are updated regularly because of the ever-changing nature of technology and the development of relevant laws in other jurisdictions and trading partners.

12.1 Classifications of Personal Information in Relation to an Individual's Public or Private Life

We can differentiate between two types of individual personal information:

1. *Personally identifiable information (PII) or sensitive personal information (SPI)*: This includes any piece of information that can, on its own or in combination with other info, uniquely or semi-uniquely identify a specific individual. Examples include full name, birth date, username on social media platforms, resume and work history, government-issued identification (passport, driving license and social insurance number), email address, telephone number, mail address, property information, communication records and content, personal picture, biometric data, credit card number, bank account number and any factor that can uniquely make a person identifiable.
2. *Anonymous information*: This type of info is not strictly related to an individual. Hence, we cannot use it solely to distinguish someone online or offline. An example of such information includes race, national origin, languages spoken, gender identity, blood type, physical traits (height, weight, age, hair color, skin tone, tattoos), income brackets, geographic location (country, GPS coordinates) and online browsing activities such as browsing behavior, links clicked, browsing history.

There was a debate about whether an internet user IP address constitutes PII or not. To answer this question, I will return to a court decision issued by *The*

European Court of Justice (ECJ), which considers internet users' IP addresses to be personally identifiable information (PII)¹³. So, to stay in the safe zone, it is better to consider an IP address as a type of PII information, although this rule has not been implemented in all jurisdictions around the globe.

Another thing we should be aware of, as is generally mentioned in most data protection laws worldwide, is the concepts of the *data controller* and *data processor*.

The *data controller* is the legal entity (individual, public authority, agency, private company) that determines on its own or in partnership with other entities the purpose of collecting and processing of consumer personal data (consumer is also known as *data subject* in most data protection laws). The controller is the entity that directs the activities of the data processor.

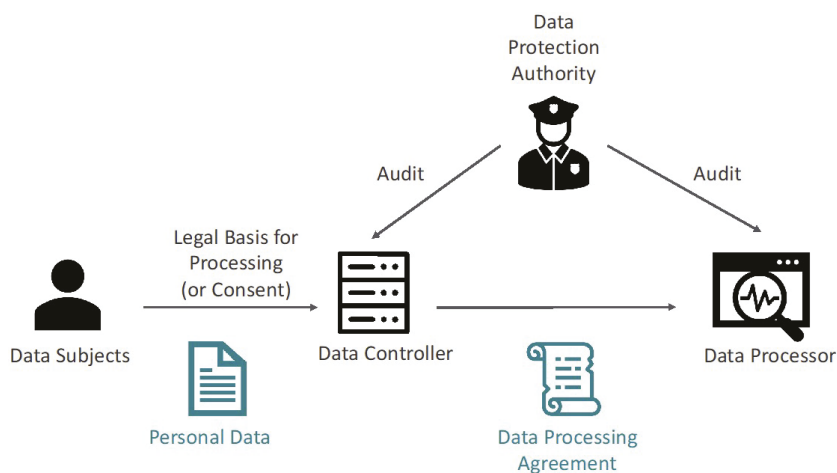
The *data processor* is the legal entity (individual, public authority, agency, private company) that processes, stores, or transmits personal data on behalf of the data controller. Data processors can only use collected data as instructed by the data controller and is commonly required to keep an audit of all processing activities.

We will give an example to clarify the concept: Most websites use third-party services to serve advertisements and to collect statistical information about their users, for instance when you visit a website (e.g., CNN website) that uses the *Google Analytics* service to analyze visitors' behaviors; the CNN website is considered the data controller while Google Analytics is the data processor. Another example is when a website uses a provider for email marketing campaigns, the original website visited by the user is the data controller, while the email marketing provider used to send emails and track user engagement is the data processor.

Data protection laws impose different obligations on the data controller and data processor. For example, under the GDPR law (see Figure 12.1), the controller is the main party responsible for consent and governing access to consumer data. It is responsible for the lawfulness, fairness, and transparency of information in addition to its responsibility for the confidentiality of personal data. The controller should select a data processor that complies with the GDPR act.

Now that we know the difference between PII and other anonymous information related to individuals and can differentiate between a data controller and a data processor, we will begin talking about the main cybersecurity and data protection regulations in the main Asia Pacific countries.

¹³Enterprise Times, "ECJ rules IP Address is PII" <https://www.enterprisetimes.co.uk/2016/10/20/ecj-rules-ip-address-is-pii> Accessed 2024-05-01

Figure 12.1: Difference between data controller, processor and data subjects under EU GDPR.

12.2 Singapore

The Personal Data Protection Commission (PDPC) in Singapore is the authority responsible for administering and enforcing the *Personal Data Protection Act (PDPA)* (<https://www.pdpc.gov.sg>). The regulation was implemented in phases. The last one came into force on 2 July 2014.

The PDPA is a general umbrella that holds many government laws concerning collecting and using individual personal data (stored in digital or non-digital forms). PDPA gives individuals the right to protect their data and governs how businesses can use personal data collected from consumers for legitimate purposes. To comply with the PDPA Act, there are different requirements that each company needs to comply with according to the industry it belongs to when collecting and processing personal data.

12.3 Japan

Directly after implementing GDPR law in the EU, Japan and the European Union agreed to recognize each other's data protection laws as providing sufficient protection for an individual's personal information. This allows enterprises working in both the EU and Japan to exchange personal information freely without any legal barrier. The framework for the mutual and easy transfer

of personal data between Japan and the European Union came into force on 12 January 2019.

The Personal Information Protection Commission (PPC) (<https://www.ppc.go.jp/en>) is an independent official authority responsible for protecting the rights and interests of individuals in the privacy and supervising the use and retention of consumer personal data by businesses. PPC is also responsible for international cooperation between Japan and other jurisdictions in the area of data protection laws.

12.4 Vietnam

In January 2019, the Vietnam cybersecurity law (http://platform.dataguidance.com/sites/default/files/vietnam_cybersecuti_law.doc) came into force. This law imposes many restrictions on domestic and foreign companies working or wanting to work in the Vietnamese market. For instance, all companies offering internet and telecommunications services or any other service related to internet or telecommunication technology (such as cloud storage providers, social networking sites like Facebook and Twitter, instant messaging services like WhatsApp, online payment systems, online merchants, domain name and hosting providers, online gaming, email providers) that operate in Vietnam cyberspace and process/retain information about Vietnamese users, must have a physical local branch or a representative in Vietnam. The law also requires such companies to store the processed data of Vietnamese users for a period specified by the Vietnamese government. The data localization element of the law is considered the most demanding part of the regulation, as it requires storing processed data in specific geographical locations within the country or handing this info to authorities; as a result, virtual companies (that operate entirely online) cannot offer their services in the Vietnam market.

It is not clear whether the Vietnamese government has the required resources, expertise and tools to enforce such strict regulations; however, we can expect to see more countries in the region move to apply similar rules to the Vietnamese government, which is somehow identical to the Chinese cybersecurity regulations that impose tight control over the internet and on all companies operating in the Chinese cyberspace.

12.5 China

In China, there are many regulations – issued by different government bodies or ministries – related to cybersecurity and internet control laws. However, in

this book, I will focus on the rules related to the protection of user personal information.

The *China Personal Information Security Specification* (<https://www.tc260.org.cn/front/postDetail.html?id=20200918200432>), which went into force in 2017, is the Chinese version of the EU GDPR and the first specification issued to protect Chinese citizens' personal data. Published by the *Standardization Administration of China*, this specification addresses the collection, transfer, and disclosure of Chinese citizens' personal information. It also defines the terms under which businesses can collect/share personal information about users, how to store and process this info in addition to required procedures to handle security incidents.

An update – or a draft measure – of this specification was issued in June 2019 and mainly addresses the transfer of important personal information across borders. The draft measures imposed the following terms on companies operating in Chinese cyberspace and handling Chinese personal information:

1. Network operators in China are required to conduct a security assessment of their systems that reveals the risks associated with the transfer of personal information outside the borders, and these assessments must be handled by the local cyberspace administration authority. This requirement raises concerns between foreign companies operating in China, as to meet this regulation companies may be required to reveal sensitive information and/or critical business secrets such as the source code of their programs/applications, critical information about their systems (e.g. encryption mechanism) to the authorities.
2. Significant data breaches should be appropriately reported to the authority without any delay. It also requires companies processing Chinese citizens' information to have an incident response plan, and conduct regular cybersecurity training of its employees, and if an incident takes place, companies should cooperate with the authorities to investigate the incident and collect related digital evidence.
3. Important personal data should be stored locally within China unless the business has passed the required security assessments imposed by official authorities, among other terms. For data affecting national security and/or having a negative effect on public interest, they cannot be transferred outside borders under any condition. Companies offering online services (e.g. WhatsApp) or other value-added services in the Chinese market must store their data locally on Chinese servers; otherwise, they are not allowed to conduct business in the Chinese market.

All companies operating in China or want to access the Chinese market should be familiar with the updated draft measures of the *China Personal Information Security Specification*; when a company cannot adhere to the draft measures' requirements (especially the security assessment part), data localization becomes mandatory to remain operational in this market.

12.6 Thailand

The Thai government released the Personal Data Protection Act (PDPA) (https://www.dataguidance.com/sites/default/files/entranslation_of_the_personal_data_protection_act_0.pdf) on 17 May 2019; the law will take effect on 27 May 2020. The Thai PDPA has extended the scope of its geographical application to include any company outside Thailand that processes or stores personal data of Thai citizens as a part of the services/products offered, regardless of whether there is a payment or not.

After reading the law, I conclude that the Thai government has adopted a similar approach to GDPR when defining the obligations of companies concerning collecting and safeguarding the personal data of individuals. The following are the main key points of the Thai PDPA:

1. There should be a legal basis for collecting the data from the consumer. In some instances, the legal basis can be explicit consent from the consumer in a written statement or via other electronic means. The consumer also should have the right to revoke access or update their data at any time and have the right to know the purpose of collecting or disclosing their personal data. Organizations should not collect personal data when they do not need it to offer the specified product/service to the consumer.
2. The act imposes a notification requirement regarding any data breach that must be executed within 72 hours after the organization becomes aware of it. The affected consumer should also be notified if the breach constitutes a high risk to their data.
3. The law requires a local representative in Thailand for some types of businesses.
4. The data controller cannot send consumer personal data outside the Thailand border without proper consent from the data owner unless the destination country has proper privacy and data protection laws or this transfer is permitted by law.
5. Breaching the PDPA law may result in severe, civil, criminal, and administrative penalties exceeding THB5m (more than US\$153,000).
6. The law allows the data controller who collects personal data of Thai consumers before enforcing this act (before 27 May 2020) to continue using it under the following two conditions:
 - 6.1. Give consumers a withdrawal method to stop using their data.
 - 6.2. If the consumer grants permission to the data controller to continue using their data, data should be used for the original purpose it was collected for and not for anything else.

Although the Thai PDPA is modeled on the EU GDPR, however, there are some key differences between both laws that make the GDPR stronger in terms of enforcing strong protection of individual data. For example, PDPA does not

explicitly set rules to control the automatic processing of personal data which is used to create a profile for internet users. PDAP also does not strictly detail the obligations of the data controller and the data processor, similar to GDPR.

12.7 Chapter Summary

Entities doing business or looking to invest in the Asia Pacific market should be aware of the different data protection and cybersecurity laws enforced by different countries in the region. Organizations should also update their legal consents – when collecting personal information from consumers – and develop privacy policies to reflect the requirements imposed by these laws. In some countries, data localization is required when your work involves collecting and retaining sensitive personal information about local consumers. Please refer to further reading below for a more comprehensive review of each country’s data protection and cyber security law.

Further Reading

1. APEC (Asia-Pacific Economic Cooperation) Privacy Framework [https://www.apec.org/Publications/2017/08/APEC-Privacy-Framework-\(2015\)](https://www.apec.org/Publications/2017/08/APEC-Privacy-Framework-(2015)) Accessed 2024-05-01
2. For more information about Singapore PDPA, please visit *The Personal Data Protection Commission (PDPC)*: <https://www.pdpc.gov.sg> Accessed 2024-05-01
3. The official PDF of the GDPR (EU) <https://gdpr-info.eu> Accessed 2024-05-01
4. Unctad, “Data Protection and Privacy Legislation Worldwide” <https://unctad.org/page/data-protection-and-privacy-legislation-worldwide> Accessed 2024-05-09
5. Dlapiperdataprotection, “DATA PROTECTION LAWS OF THE WORLD” <https://www.dlapiperdataprotection.com> Accessed 2024-05-09

Index

C

cyber threat intelligence (CTI) 1, 3, 13, 14

D

data analysis 11, 72
digital footprint 13, 24

G

Geolocation 67

O

online investigations 13, 21, 33, 86
open source intelligence (OSINT) 1, 12, 13, 15, 16, 21, 43, 93, 106
OSINT tools 1, 67, 95

P

privacy and anonymity 33
profiling 23, 41

S

social media intelligence (SOCMINT) 13, 55, 63, 72

T

threat intelligence 1, 3, 4, 6, 7, 8, 10, 11, 12, 13, 14

W

web scraping 58, 72